

THE Public Interest

NUMBER 98, WINTER 1990, PRICE: \$4.50

Social Science and Black/White

Still an American Dilemma
R. J. Herrnstein

Social Psychology's "Racism"
Byron M. Roth

✓ Employment Testing and Job Performance
Jan H. Blits and Linda S. Gottfredson

Tissue Transplants: The Dilemma
of the Body's Growing Value

Emanuel D. Thorne

Rediscovering the
Three-Decker House

Howard Husock

The Work Ethic—Then and Now

Seymour Martin Lipset

The "Japanization" of America?

Steven Kelman

The Myth of the
General Right to Bail

Robert F. Nagel

Robert Bork and the
Crime of Inequality

Aaron Wildavsky

Politics, Nonpartisanship,
and the Public Service

Edward J. Lynch

Employment testing and job performance

JAN H. BLITS & LINDA S. GOTTFREDSON

FOR MORE than twenty years, a fierce public-policy storm has raged over how or whether general mental ability tests should be used in employment selection. Employers typically favor using such tests because they are generally the best single predictor of job performance. But the tests are controversial because they disproportionately screen out blacks, who as a group tend to score considerably lower than whites.

Even though the 1964 Civil Rights Act specifically sanctions general ability tests, the federal government sharply curtailed their use, ostensibly to comply with Title VII of the Act. In 1970 the Equal Employment Opportunity Commission (EEOC) issued guidelines that the Supreme Court soon adopted as law in the landmark case of *Griggs v. Duke Power Company* (1971); after *Griggs*, any employment test that had an adverse impact on blacks was considered discriminatory and therefore illegal unless it could be shown to be specifically job related. In the wake of subsequent court decisions and stringent federal regulations for demonstrating

job relatedness, employers virtually ceased using general mental ability tests.

In 1981, however, as a result of new research on testing, the Labor Department's Employment Service began encouraging state employment services to use its General Aptitude Test Battery (GATB) for all job referrals and to rank candidates according to test scores. Intended to improve the state agencies' screening of nearly 20 million applicants a year, the Department's action was based on psychological research that undercut the EEOC's original criticism of general ability tests. The original criticism rested on the fact that a given test's ability to predict job performance—its "validity"—seemed to vary from job to job, and from location to location. Because a test that was valid in one location often appeared to be invalid in other locations for the same job, experts tended to believe that every test needed to be validated separately for each new setting.

The new research, pioneered largely by Frank Schmidt and John Hunter, challenged the need for separate validation. By applying new techniques, Schmidt, Hunter, and other researchers showed that well-developed mental ability tests can be generalized to new locations and even to new jobs. Most of the previously observed variation in a given test's ability to predict performance from one job or setting to another was shown to be merely a statistical artifact caused by small sample sizes.

This new research has disconcerted many opponents of testing, because, while showing that general ability tests are valid for most skilled and semiskilled jobs, it does nothing to reduce their adverse impact. In fact, it shows that the persistent adverse impact of these tests cannot be blamed on the inadequacies of the tests themselves. Several decades of other psychological research have also refuted the still popularly held belief that general ability tests are racially biased—that is, that they underpredict the job performance of blacks. Many separate studies show that, contrary to prevailing public opinion, lower test scores among blacks and Hispanics are accompanied by lower job performance, just as in the case of whites.

The new research has thus revived and intensified the policy storm over general mental ability tests; the question now is not whether but how such tests should be used. Specifically, can they be used without incurring adverse impact?

The Labor Department found a way. Its 1981 action promoted

not only the use of the GATB to improve productivity, but also the adoption of a race-conscious way of recomputing test scores to avoid adverse impact. The new procedure, called race-norming—or, euphemistically, “within-group scoring”—converts raw scores into group-based percentile ranks. Each candidate’s score is reported not in relation to those of all other candidates, but only in relation to the scores of applicants of the same group. Blacks are compared only to other blacks, Hispanics only to other Hispanics, and “others” to all but blacks and Hispanics. Top scorers from all groups are then referred on the basis of their group-based percentiles to employers, who are often not aware that the candidates’ scores have been so adjusted. In effect, race-norming gives bonus points to members of groups that tend to score lower than others, offsetting average differences in scores among groups. By 1986 state employment agencies in forty states had adopted the Labor Department’s new referral procedure.

Race-norming crucially affects employment referrals. Among candidates for semiskilled jobs, for example, blacks at the 50th percentile on the GATB have raw scores of 276; Hispanics, 295; and whites, 305. A score of 305 is at the 84th percentile among blacks. Thus the raw score that would place a candidate at the 50th percentile among whites would place him at the 84th percentile among blacks. Likewise, among candidates for many skilled jobs, a black applicant with a score of 283 would have the same ranking and referral status (70th percentile) as a white applicant with a score of 327.

Faced with a 1986 challenge from the Justice Department (which held that race-norming constitutes unconstitutional reverse discrimination), the Labor Department agreed not to extend its program until the National Research Council of the National Academy of Sciences (NAS) conducted a thorough scientific evaluation of it. The NAS panel recently issued its report, *Fairness in Employment Testing: Validity Generalization, Minority Issues, and the General Aptitude Test Battery*;¹ this report has inadvertently provoked other, even more troublesome questions concerning not the impact of tests on racial groups, but the impact of politics on science. What is the proper role of political views in formulating important “scientific” evaluations? Is science being used as a guise for purely personal political preferences?

¹John A. Hartigan and Alexandra K. Wigdor, eds. National Academy Press. 334 pp. \$29.95 (paper); \$39.95 (cloth).

The NAS report

In its report, the NAS panel, headed by Yale statistician John Hartigan, confirms that the GATB’s validity for predicting job performance is strong enough to enhance worker productivity; but it recommends that the tests not be used unless the results are race-normed. The panel was unable to justify race-norming on the grounds of racial bias—it confirms that tests like the GATB may even slightly overpredict the performance of blacks, particularly in higher level jobs; in this sense the tests are slightly biased in favor of blacks. Nor did it find that the tests were irrelevant—the panel found a 0.3 correlation between test scores and job performance, which means that employers can realize 30 percent of the gains in workforce productivity that a perfectly valid test would yield. Undeterred, the panel looked elsewhere and found—or at least claimed to find—justification for race-norming in the very nature of a selection system that favors people with high scores over those with low scores.

Any selection system that selects higher scorers over lower scorers necessarily favors those who are above a cut-off rather than those who are below it. However, some candidates who score below a cut-off may in fact be able to perform the job at least at a minimally acceptable level; such candidates are called “false negatives.” Others, “false positives,” score above the cutoff but are nonetheless unable to perform at a minimally acceptable level. The lower a test’s validity, the more such false predictions are likely to occur. Now, more blacks than whites fall in the category of “false negatives,” and more whites than blacks are “false positives.” The reason for these disproportions has nothing to do with race *per se*, but arises from the fact that more whites have scores above the cut-off and more blacks below it. “[T]hese effects,” the panel agrees, “are a function of high and low test scores, not racial or ethnic identity.” Yet even as it emphasizes that the effect is the same for all low-scoring individuals, regardless of their race, the panel nevertheless claims that “the disproportionate impact of selection error provides scientific grounds for the adjustment of minority scores.” It concludes, in other words, that the disproportions in false predictions provide a “scientific justification” for benefiting test takers whose scores—if anything—overpredict their job performance.

The central failing of the NAS panel is that it recommends a race-based solution to what it concedes is not a race-based problem;

thus it seeks a scientific justification for what is scientifically unjustifiable. The panel evidently set out to endorse race-norming; to do so, it had to make slanted analytic assumptions, to choose a questionable criterion for determining the tests' fairness, to disregard technical distinctions, and to frame policy considerations one-sidedly.

For example, the panel consistently interprets its data in ways that minimize the GATB's validity. Repeatedly choosing what it acknowledges to be the most "conservative analytic assumptions," it is able to limit the tests' estimated validity to 0.3. (Labor Department technical reports place it at 0.5 or higher.) Time and again the panel then belittles this correlation as "imperfect" or "only modest," despite admitting that it is strong enough to increase productivity. Since the tests' validity poses a stumbling block to any attempt to offer "scientific justifications" for race-norming, the panel is forced to disparage it. The better a test predicts job performance, the more "deceptive" (the panel's word) race-norming becomes.

Even as the panel chooses analytic assumptions that limit the tests' validity, it also sets a standard of perfect validity. The panel shows that the unadjusted GATB scores meet the only widely accepted criterion of fairness in testing—that a test predict equally well for all groups; it then drops that criterion, however, and adopts another—that the test's false predictions not adversely affect minorities—that the unadjusted scores fail to satisfy. In fact, though, any test that meets the commonly accepted criterion must necessarily fail to meet the panel's criterion. (Conversely, any test that meets the panel's criterion must necessarily fail to meet the commonly accepted one.) Only a perfectly valid test could meet both criteria; and perfectly valid tests do not exist. Most experts reject the panel's chosen criterion, viewing it as internally inconsistent; but the panel adopts it without reviewing the scientific literature critical of it or even acknowledging that serious criticism exists.

The panel also draws attention only to the crudest distinctions in job performance, ignoring differences between minimally competent and highly productive workers. For example, when asserting the unfairness of using unnormed test scores, the panel indiscriminately groups together as "able" all workers whose job performance is above a minimum level. This procedure has the effect of exaggerating the significance of false predictions, providing spe-

rious strength to the panel's "scientific justification" for race-norming. Similarly, the panel recommends reporting a second, unnormed score for all candidates already referred on the basis of race-norming; the second score would indicate the probability that a candidate will perform "above average." Although ostensibly meant to reveal intergroup distinctions that race-norming (admittedly) conceals, the unnormed score in fact conceals the full extent of differences in expected job performance because it makes no distinctions among higher levels of performance. Its sole standard is "above average" performance.

Policy considerations

The panel's discussion of policy considerations is similarly one-sided. The panel warns, for example, that the renewed popularity of mental ability tests could lead some groups to feel superior and others to feel inferior; but it says nothing about the feelings that race-norming might produce among those same groups. In discussing race-conscious selection procedures, it cites the opinion of the Lawyers' Committee for Civil Rights Under Law, but not that of any of the group's mainstream or conservative counterparts. It repeatedly warns that the adverse impact of general ability tests makes employers vulnerable to Title VII discrimination suits, but it never points out that its own findings demonstrating the validity of the tests may constitute an employer's defense. It frequently emphasizes the possible advantages of race-normed referrals, but it never clearly assesses the economic costs or legal risks to employers. Although it emphasizes that the problem of false predictions is a function of low scores and not of race, it nonetheless recommends adjusting scores to benefit only minority "false negatives," not white ones.

Further, the panel misconstrues the obvious meaning of the phrase "adversely affect" in Title VII of the 1964 Civil Rights Act, so that it can claim that the title's definition of discrimination centers on unequal results for groups rather than on unequal opportunities for individuals. Contrary to the very words it cites, the panel concludes that "Title VII ... adopts a group-centered definition of discrimination, outlawing 'employment practices' that 'adversely affect' an individual's status as an employee because of that employee's race, color, religion, sex, or national origin." More generally, almost whenever the panel speaks of "equality of opportunity," it really means equality of results for groups.

Likewise, when the panel discusses equal-protection jurisprudence and affirmative action, it relies on the constitutional interpretation of only one scholar, Laurence Tribe. And when it presents the arguments for and against preferential treatment, it inverts the natural order, placing the argument for after the argument against, and then concludes with a section devoted entirely to discrediting the case against preferential treatment.

In that section (entitled "Beyond Philosophy"), the panel, evidently speaking for itself, provides two arguments "to cut through the intellectualization of the issue of preferential consideration for blacks or other disadvantaged minorities." It first contends that preferential treatment has always existed in America, with only the intended beneficiaries having changed. "Very powerful forces support preferential treatment for veterans, including hiring preference in the civil service and referral priority by the U.S. Employment Service." But preferential consideration on the basis of race differs from grateful repayment for service to the nation; more generally, granting people rights merely because of their race differs from grouping together individuals with the same preexisting rights and claims.

The panel's second argument asserts "a skeptical assessment of the liberal values of equality, color-blind law, merit, and fair competition," as seen by those barred from enjoying them until the passage of the 1964 Civil Rights Act. In their view, the panel claims, domination over blacks was sanctioned first by religion (in the colonial period) and then by Social Darwinism (after the Civil War); now it is buttressed by "the myth of equality," which "provides a veneer for further oppression." If the panel is to be believed, just as preferential treatment is nothing novel in America, so too equality of opportunity is only a twentieth-century mask to hide racial oppression.

Politicizing science

Early in its report, the panel observes that "[i]n hindsight it is clear that many of the advocates of early [ability] testing allowed their scientific judgment to be influenced by contemporary racial and ethnic biases and by unexamined assumptions about the social order". The same is no doubt true of the panel itself, whose "scientific judgment" is clearly influenced by policy preferences. Race-norming seems to have been the panel's intended solution from the start.

What the panel in fact shows is not that race-norming is scientifically justified, but that the decision to race-norm is not a scientific issue. The panel confirms that testing's adverse impact is not caused by defects in the tests themselves. Hence the adverse impact has no technical solution, let alone one that requires adjusting minority scores to compensate for "the inadequacies of the technology [of testing]." Nor is race-norming an issue on which scientists are in a special position to make policy recommendations. Instead, race-norming is a political question that ought to be answered politically. Whether the nation should adopt race-norming should be decided by public officials and ultimately by the citizens; to pretend that science can supply an answer is only to pervert science by politicizing it.