The Science and Politics of Race-Norming

Linda S. Gottfredson

Disparate impact *(racial imbalance) in employee selection constitutes prima facie evidence of unlawful discrimination. Research in personnel psychology has shown, however, that valid and unbiased selection procedures often guarantee disparate impact and that they will continue to do so as long as there remain large racial disparities in job-related skills and abilities. Employers are in a legal bind because often they can avoid disparate impact only by engaging in unlawful disparate treatment (racial preferences). Some personnel psychologists have argued that there is scientific justification for race-based adjustments in test scores that eliminate disparate impact. Analyses of their seemingly scientific reasoning illustrate how personnel selection science is being compromised in an effort to reconcile contradictory legal demands.*

In 1989 a blue-ribbon committee of the National Academy of Sciences (NAS) concluded that the U.S. Department of Labor's practice of race-norming employment test scores was "scientifically justified" because "the modest validities of the GATB [General Aptitude Test Battery] cause selection errors that weigh more heavily on minority workers than on majority workers" (Hartigan & Wigdor, 1989, p. 7). In just two years, however, all forms of race-norming, including the Department of Labor's within-group score adjustments, would become illegal. With strong public support and facing virtually no opposition, both Democrats and Republicans in Congress joined in adding a ban on such race-conscious score adjustments to the Civil Rights Act of 1991.

How could an employment practice with such an apparently compelling rationale and authoritative support be outlawed? Was this just another case of politics trumping scientific knowledge?

Legal Constraints Versus Selection Realities

In hindsight, race-norming (a species of subgroup norming) was the logical outcome of growing legal and regulatory constraints in personnel testing. Title VII of the Civil Rights Act of 1964 outlawed disparate *treatment* by race, meaning, for example, that Blacks could no longer be refused employment or evaluated differently on the basis of race. By 1970, the Office of Federal Contract Compliance Programs had issued regulations, based on then-President Lyndon Johnson's 1965 Executive Order No. 11246, requiring all federal contractors to implement written, result-oriented, affirmative action compliance programs with specific goals and timetables for minority hiring. In 1971, the U.S. Supreme Court gave further impetus to such results-driven civil rights enforcement policy when it ruled in *Griggs v. Duke Power Co.* that disparate effects constitute prima facie evidence of unlawful employment discrimination. Even objective, unbiased, professionally developed tests were now unlawful if (a) plaintiffs could demonstrate they resulted in racial differences in selection (disparate impact), and (b) the employer failed to prove that those tests were required by "business necessity," that is, they were job related (valid) for the job in question.

Although *Griggs v. Duke Power Co.* virtually paralyzed employment testing in the short term, it helped force personnel psychology to turn more concerted attention to test bias and criterion-related validity. Researchers hoped, if not expected, that they could identify or develop valid, unbiased tests with less (or perhaps no) disparate impact. With more valid selection procedures, "nondiscriminatory treatment" would lead in time to "nondiscriminatory results."

Unfortunately, research and experience in the intervening decades have revealed that disparate impact is the rule, not the exception, even when using valid, unbiased tests—especially cognitive tests, which are generally the best predictors of job performance (e.g., Schmidt & Hunter, 1981). The disparate impact of such tests is due not to their imperfections but to substantial racial differences in the job-related skills, abilities, and knowledges they reveal (Wigdor & Garner, 1982). Their use cannot be abandoned without considerable sacrifice in workforce productivity.

Personnel psychology has thus learned a perverse truth in employment testing: The disparate impact and disparate treatment standards will create mutually contradictory demands as long as racial–ethnic groups continue to differ substantially in their job-related skills and abilities (e.g., Gottfredson, 1988; Schmidt, Ones, & Hunter, 1992; Sharf, 1988; Wigdor & Sackett, 1993). The

Editor's note. At the request of the action editor, the focus of this article is the controversy that led to the amendment of the Civil Rights Act of 1991 and racial–ethnic subgroup norms for tests of cognitive ability. The article does not address all of the issues raised by Sackett and Wilks (1994, this issue); rather, it provides a penetrating analysis and contrasting view of fairness in employment testing.

Author's note. The author gratefully acknowledges the helpful comments on this article made by Jan H. Blits, Robert A. Gordon, Frank L. Schmidt, and James C. Sharf.

Correspondence concerning this article should be addressed to Linda S. Gottfredson, College of Education, University of Delaware, Newark, DE 19716.

 0003-066X/94/$2.00
Vol. 49, No. 11, 955-963

quandary for employers is that, on the one hand, using merit-based, color-blind tests generally guarantees disparate impact and thus invites litigation under the disparate impact standard of discrimination. On the other hand, using racial preferences is often the only feasible way for an employer to eliminate adverse impact in the short run, but such preferences constitute discrimination against nonminorities and thus invite litigation under the disparate treatment standard.

Pragmatism at the U.S. Employment Service

The U.S. Employment Service (USES) of the Department of Labor provides a good case study of where the disparate impact standard has been leading personnel psychology. As more fully described elsewhere (Hartigan & Wigdor, 1989), it led USES to make a fateful choice about its GATB in 1980.

The state employment service offices to which USES provides technical assistance had never made much use of its decades-old GATB in selecting job applicants for referral to participating employers. Their use was minimal because USES scored the GATB in a way (multiple low-hurdle cutoffs) that retained little of its predictive value and because the GATB's use had been validated for only a few hundred of the thousands of jobs for which the state employment service offices make referrals.

By 1980, extensive research on the GATB, much of it using the powerful new techniques of meta-analysis, had led USES to conclude that, when scored and used in an optimum manner, the GATB is useful in predicting performance in virtually all jobs, that it is not biased against Blacks or Hispanics, and that its greater use by employers would substantially improve workforce productivity in the United States. However, the GATB had one feature that USES feared would deter most employers from using it—severe disparate impact. Blacks score about one standard deviation and Hispanics about one-half standard deviation below Whites on the GATB, levels of disparate impact that we now know are fairly typical of valid employment tests (Wigdor & Garner, 1982). USES was, no doubt, well aware that employers, when put in the untenable position of having to violate one standard of discrimination in order to satisfy the other, generally opt to treat applicants differently on the basis of race in order to avoid racial imbalance in hiring.

USES therefore decided to eliminate the GATB's disparate impact by race-norming applicants' test scores. Specifically, it produced separate percentile conversion tables for Blacks, Hispanics, and others (primarily Whites and Asians) by which the state employment service offices could convert job applicants' raw GATB scores into percentile rankings within each racial group (within-group scores) for each of five different job families. For example, among candidates for many skilled jobs (Job Family I), a GATB raw score of 300 translated into percentile scores of 79, 62, and 38, respectively, for Blacks, Hispanics, and others.

Job applicants were then referred to employers on the basis of their race-adjusted percentile scores, a fact that applicants never knew and few employers ever understood if told. These percentile conversions allowed the state employment service offices to refer equal proportions of applicants from each racial group, despite substantial racial gaps in job-related skills and abilities. Typically, for example, a White or Asian person would have to score around the 84th percentile to have the same chances of referral as a Black person scoring at only the 50th percentile for Whites.

USES knew that its race-based score adjustments would reduce the productivity gains that the unnormed GATB made possible (Hunter, 1983). USES reasoned, however, that national productivity would suffer less if employers made extensive use of a race-normed GATB in order to racially balance their workforces than if they continued to use methods that more seriously damaged productivity. When one opts for racial parity in hiring, despite large racial disparities in skill and ability among applicants, within-group scoring is the race-conscious procedure that degrades predictive validity the least.

When the U.S. Department of Justice eventually became aware of the USES practice, it threatened to sue USES for reverse discrimination, because within-group scoring is clearly a quota system. Its purpose and its effect are solely to produce racial parity. USES and the Justice Department reached an agreement in 1986 by which USES would cease expanding application of its new scoring system until a NAS committee, to be convened, rendered a verdict on the technical quality of the GATB and the scientific merits of USES's new system for scoring it.

National Academy of Sciences Committee's Reasoning on Race-Norming

In its final report (Hartigan & Wigdor, 1989), the NAS committee criticized certain features of the GATB and argued that USES had overstated its utility. However, it concurred with USES that the battery is not biased against Blacks and Hispanics; that, in fact, like most other selection tests, it slightly *over*predicts Blacks' job performance (mostly a statistical artifact due to less-than-perfect reliability); that it is valid for predicting performance in most jobs USES handles; and that its validity and utility are substantial enough for individual employers to benefit from its use.

The committee thereby seemed to have provided evidence that the GATB could pass legal muster. It nonetheless concluded that some form of race-conscious score adjustments was necessary to avoid unfairness to lower scoring minority groups.

> . . . Majority workers do comparatively better on the test than they do on the job, and so benefit from errors of false acceptance. Minority workers at a given level of job performance have much less chance of being selected than majority workers at the same level of job performance, and thus are burdened with higher false-rejection rates. . . . This outcome is at odds with the na-

tion's express commitment to equal employment opportunity for minority workers. In the committee's judgment, the disproportionate impact of selection error [on minority workers] provides scientific grounds for the adjustment of minority scores. (Hartigan & Wigdor, 1989, p. 7).

The remedy, the committee suggested, is therefore to adjust minority scores upward until "able minority workers have approximately the same chances of referral as able majority workers" (Hartigan & Wigdor, 1989, p. 7), for example, via within-group scoring (as USES had done) or with the committee's own variety of race-norming (which it labeled *performance-based* adjustments). The former always eliminates the entire racial gap in reported test scores; the latter eliminates somewhat less of the gap when a test is more valid. The less valid the test, the larger the adjustments must be. In the simplest case, performance-based score adjustments reduce a racial gap (m) in test performance by a factor of $1 - r_{xy}^2$ (Hartigan & Wigdor, 1989, p. 262).

So, for example, adjustments for a test having a validity coefficient (r_{xy}) of .3 would reduce a mean racial difference by 91%, thus turning the common 1.0 *SD* mean Black–White difference in unadjusted scores into an 0.09 *SD* difference in adjusted scores. The NAS committee estimated that most GATB true validities range between .2 and .4, meaning that "performance-fair" adjustments would typically eliminate from 84% to 96% of any racial gap. A performance-fair system would thus calibrate its adjustments to a test's level of predictive validity, but the need to have a precise estimate of validity would also make the system more difficult to implement than would within-group scoring. In most circumstances, both forms of race-norming would produce "virtually identical" (Hartigan & Wigdor, 1989, p. 272) results, so the NAS committee expressed a "slight preference" for the simpler form, namely, USES's "within-group scoring" (p. 271).

Public Reactions

Public responses among personnel researchers to the NAS committee's recommendation on subgroup norming were mostly negative and unusually biting. Most critics (Humphreys, 1989; Schmidt, 1990; Tenopyr, 1990) protested that the committee had tried to cloak a political preference in questionable science. Some (Blits & Gottfredson, 1990a, 1990b) argued, in addition, that race-norming is destructive social policy because, among other side effects, it would make permanent the very social inequalities it is supposedly intended to eliminate. Various committee members (Sherwood, 1990; Wigdor & Hartigan, 1990; Wigdor & Sackett, 1993) tried to rebut or impugn the critics, but to little avail.

Two themes among the critics were that the committee had used a highly defective model of test fairness and that it had distorted or obscured evidence in making its case. For example, the NAS committee failed to note the large literature evaluating many different models of test fairness or that its conception of performance fairness (based on the Cole–Darlington conditional probability model of selection fairness) had been discredited years earlier (Hunter & Schmidt, 1976; Petersen & Novick, 1976). As several critics pointed out, the model is internally inconsistent (leads to logically contradictory conclusions) and produces very different estimates of predictive unfairness, depending on where pass–fail cutoffs are set.

The model also ignores valid distinctions in test and job performance by collapsing both into mere pass–fail dichotomies when calculating rates of prediction error. Such collapse of meaningful differences was essential to sustaining the committee's claim that all "good" workers (from the minimally acceptable to the truly exceptional) are "equally able" workers, which in turn was critical to the committee's rationale for race-based score adjustments (that good Black workers have the same chance of referral as good White workers). In fact, as the committee itself implicitly conceded (Hartigan & Wigdor, 1989, pp. 264–265), good White workers tend to outperform good Black workers.

Tenopyr (1990) pointed out that the 1985 *Standards for Educational and Psychological Testing* endorse the regression model of test bias (which the GATB passes but the committee's model violates). Nowhere, however, do the standards support the NAS committee's model of test fairness (which the GATB fails). Indeed, they eschew all judgments about fairness, noting that measurement bias is a technical issue on which there is expert consensus, but fairness is a social and political issue. (The NAS committee itself had stated, however, that the accepted professional definition of test bias is also the "classical" conception of test fairness and the one "most widely accepted in the psychometric literature, at least as a minimum requirement" [Hartigan & Wigdor, 1989, p. 255]).

By definition, all unbiased tests (except perfectly valid ones) fail the NAS committee's standard of fairness when subgroups differ in test performance (as they usually do). Conversely, only tests that are statistically biased in favor of lower scoring groups would ever pass the committee's new standard of test fairness (Hartigan & Wigdor, 1989, p. 255). By the committee's logic, then, test bias against Whites and Asians is a necessary component of selection fairness for Blacks and Hispanics. Indeed, the function of race-norming is precisely to introduce such racial bias in measuring job-related skills and abilities.

Several critics also argued that the committee's proposed remedy (race-conscious score adjustments) did not fit the alleged unfairness the committee sought to eliminate (race-neutral errors of prediction). Low scorers of all races are subject to such errors, as the committee repeatedly pointed out, but scores would be adjusted for every Black and Hispanic, and no others, no matter what any individual's test score. Moreover, the Blacks and Hispanics who would benefit most often from the score adjustments would be the higher scoring ones, not the lower scoring ones (the falsely rejected), in whose name those adjustments were justified, because all high scorers are selected before any low scorers in a top–down selection system. And the higher scoring Blacks would benefit more than the higher scoring Hispanics, because Blacks would

receive adjustments twice as large as those for Hispanics (because Blacks as a group tend to score twice as far below Whites as Hispanics).

Blits and Gottfredson (1990a) described the NAS committee's apparent skewing of its statistical procedures when estimating the GATB's true validities: It chose an unrealistically high estimate of criterion reliability for disattenuating the GATB's observed validity coefficients for different jobs, and it opted not to correct them at all for range restriction in applicants' test scores. (Pearlman, 1994, has described similar skewing for other analyses in the committee's final report.) The committee thus minimized the GATB's apparent validity and thereby also magnified the GATB's imperfections and the size of the required performance-fair score adjustments. (Schmidt et al., 1992, and Sackett & Ostgaard, 1994, later confirmed that the committee's procedures greatly understate true validities.)

Despite the criticisms, many employers and personnel-selection specialists were privately pleased by the NAS committee's seemingly scientific green light for race-based score adjustments. Nor was it long before staff at the Equal Employment Opportunity Commission (unbeknownst to its chairman) started using the committee's rationale to ratchet up the regulatory pressure for racial balance by threatening to sue several Fortune 500 companies unless they began to race-norm their employment tests (Gottfredson, 1990). The committee's definition of test fairness also found its way into the formal legislative history of the 1990 civil rights bill, which, had it been enacted, might have been interpreted by judges to mandate race-norming.

However, that near coup for race-norming became its death knell when the fairness language accompanying the just-vetoed 1990 civil rights bill came to public attention. Already struggling to deny that their 1991 civil rights bill was a quota bill, its embarrassed sponsors acceded to pressure to ban score adjustments on the basis of race, gender, and ethnicity. The NAS committee's seemingly compelling rationale for race-norming quickly receded into oblivion as public outrage escalated over the bald racial preferences the practice actually seemed to entail.

Politically Selective "Science"

The NAS committee, wittingly or not, had attempted to usurp a fundamentally political decision (whether to use racial preferences) by transmuting it into a seemingly technical, scientific issue over which scientists could claim special authority. Although creating the appearance of scientific logic, the committee's appeal to disproportionate false negatives among able minority workers is but a technical pretext for procedures the committee acknowledged are race conscious.

The committee's rationale capitalizes on personnel-selection, psychology's principal concern—namely, to minimize errors in predicting applicants' job performance (in the least squares regression sense), which in turn maximizes expected job performance. However, the committee's rationale actually subverts that concern by turning it to the purpose of increasing error for lower scoring groups in the name of fairness. As we shall see, the committee does so by selectively emphasizing only those kinds of error that seem to support its rationale and largely ignoring those that undermine it.

The committee's final report introduced its performance-based model of fairness as one of two "perspectives" on fairness, the other of which it called *predictive fairness.* In fact, the literature on test fairness has examined 11 formal models of fair selection with an unbiased predictor (see Jensen, 1980, chap. 9, for an overview), 10 of which require some decrement in expected criterion performance in order to achieve some other social purpose (accounting for their technical designation as quota models). Although the committee made no mention of the fact, its two perspectives actually invoke 4 of the quota models, as discussed later: the conditional probability, converse conditional probability, equal probability, and converse equal probability models (Jensen, pp. 402–404).

The committee used a concrete example to explain its preference for the performance-fair perspective, which involved data from a GATB validity study of 91 White and 45 Black carpenters. A closer look at this example reveals the committee's selective attention to error. Figure 1 sets the stage for that example by showing the conceptualization of error underlying the committee's two perspectives on fairness and the four formal models they encompass.

The ellipses in Figure 1 represent the distribution of individuals' performance both on a test and a job performance criterion, one ellipse for Blacks and one for

Figure 1
Categories of Hits and Misses in Predicting Success (High–Low) on the Criterion From Success (Pass–Fail) on the Test

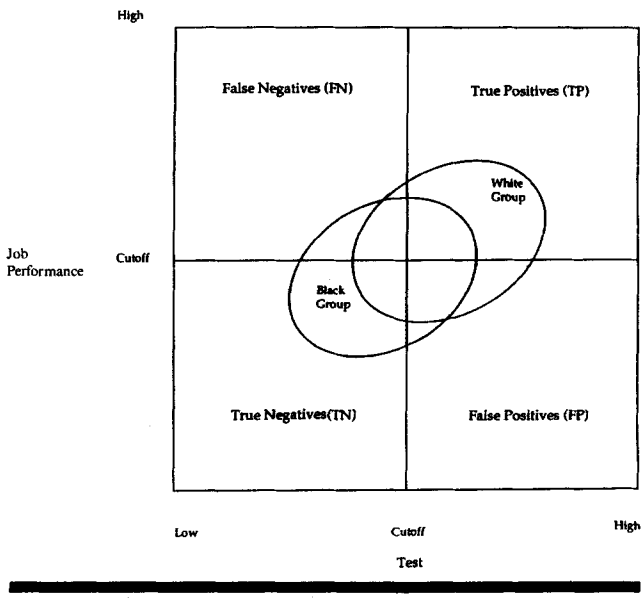

Table 1
Distribution (No.) of Blacks and Whites According to Performance on the Test Versus on the Job

	Test performance					
	Whites (N = 91)[a]		Blacks (N = 45)[a]		Blacks' adjusted scores[b]	
Job performance	Fail	Pass	Fail	Pass	Fail	Pass
Good	11	**60**	8	**8**	2	**14**
Poor	**11**	9	**24**	5	**12**	17

[a]Unadjusted test scores. [b]Scores adjusted to satisfy conditional probability model of selection fairness. Black–White unadjusted mean test score difference is about 1.25 *SD*, as calculated from $z = -.70$ for a 24% White fail rate and $z = .55$ for a 71% Black fail rate. Black good workers are reclassified to simulate adjustments that satisfy the conditional probability model. Reclassification of Black poor workers is consistent with $r_{xy} = .4$ and mimics a 1.05 *SD* score adjustment upward for all Blacks (because *z* goes from a .55 to −.50 when Black test failure rate drops from 71% to 31%).

Whites. The foregoing four models of selection fairness collapse the two continua of performance into two dichotomies by establishing arbitrary cutoffs for passing the test (the vertical line bisecting the ellipses) and for succeeding on the job (the horizontal line bisecting the ellipses). The resulting four quadrants represent two kinds of error and two kinds of success in predicting job performance from test scores: false negatives (FN, individuals who fail the test but succeed on the job), false positives (FP, individuals who pass the test but fail on the job), true negatives (TN, individuals who fail on both the test and the job), and true positives (TP, individuals who succeed on both the test and the job). As Figure 1 shows, the individuals within any quadrant often differ greatly in either their test scores or their job performance, but the foregoing models consider them all identical for purposes of assessing selection fairness.

Figure 1 also represents the typical situation in selection, namely, that tests predict performance equally well for Blacks and Whites (the regression line is essentially the same for both groups), but that many more Blacks than Whites fail the tests. The frequencies in Table 1 for Whites and Blacks illustrate these facts (ignore for now the columns for Blacks' adjusted scores). Around three fourths of both Blacks (32 of 45) and Whites (71 of 91) are correctly classified (as true negatives or true positives) by unadjusted test scores; Blacks fail the test at a much higher rate (32 of 45, or 71%) than do Whites (22 of 91, or 24%); and most Blacks fail both the test and the job (24 of 45, or 53% true negatives), whereas most Whites succeed on both (60 of 91, or 66% true positives).

The committee's two perspectives on fairness are really two ways of calculating error rates from Table 1. As shown in Table 2, the committee's preferred performance-fair perspective uses row percentages to gauge whether the two error rates differ by race. That is, the denominator for the rate of false negatives is all good workers (the first row), and the denominator for the rate of false positives is all poor workers (the second row). (Rates of correct prediction are shown in boldface type.) In effect, this perspective is concerned with error in using *job* performance to predict *test* performance.

By this perspective, both error rates seem to disadvantage Blacks. As shown in the first row, 15% of White good workers but 50% of Black good workers fail the test (are false negatives) and would have mistakenly been denied employment. In addition (the second row), relatively more White than Black poor workers passed the test (45% vs. 17% false positives) and would have been mistakenly hired.

These differences in performance-based error rates are the heart of the NAS committee's claim that prediction errors in a race-neutral selection system disadvantage minority workers who are equally able as White workers. The committee noted that these differences in error rates are race-neutral in origin and occur simply because Blacks tend to score lower than Whites on valid tests. White low scorers are just as subject to prediction errors as are Black

Table 2
Row Percentages for the National Academy of Sciences Committee's Performance-Based Perspective on Selection Fairness

	Test performance					
	Whites[a]		Blacks[a]		Blacks' adjusted scores[b]	
Job performance	Fail (%)	Pass (%)	Fail (%)	Pass (%)	Fail (%)	Pass (%)
Good[c]	15	**85** (100%)	50	**50** (100%)	13	**87** (100%)
Poor[d]	**55**	45 (100%)	**83**	17 (100%)	**41**	59 (100%)

[a]Unadjusted test scores. [b]Adjustments to satisfy conditional probability model of fairness. [c]Conditional probability model of fairness requires equal rates of true positives, TP/(FN + TP), for all races. Same as requiring equal rates of false negatives, FN/(FN + TP). [d]Converse conditional probability model of fairness requires equal rates of true negatives, TN/(TN + FP), for all races. Same as requiring equal rates of false positives, FP/(TN + FP).

Table 3
Column Percentages for the National Academy of Sciences Committee's Prediction-Based Perspective on Selection Fairness

	Test performance					
	Whites[a]		Blacks[a]		Blacks' adjusted scores[b]	
Job performance	Fail (%)[c]	Pass (%)[d]	Fail (%)[c]	Pass (%)[d]	Fail (%)[c]	Pass (%)[d]
Good	50	**87**	25	**62**	14	**45**
Poor	**50**	13	**75**	38	**86**	55
	(100%)	(100%)	(100%)	(100%)	(100%)	(100%)

[a]Unadjusted test scores. [b]Adjustments to satisfy conditional probability model of fairness. [c]Converse equal probability model of fairness requires equal rates of true negatives, TN/(FN + TN), for all races. Same as requiring equal rates of false negatives, FN/(FN + TN). [d]Equal probability model of fairness requires equal rates of true positives, TP/(TP + FP), for all races. Same as requiring equal rates of false positives, FP/(TN + FP).

low scorers. The committee argued, however, that relatively more Blacks score low, and thus more are subject to what it called damaging errors (false negatives).

Table 3 shows the column percentages that the committee used to illustrate its second, prediction-based perspective on fairness. The denominator for the false-negative rate is now all applicants who fail the test (the first column for both races), and the denominator for the false-positive rate becomes all applicants who pass the test (the second column). In effect, the prediction-based perspective is concerned with error in using *test* scores to predict *job* performance (which is the normal concern in personnel selection).

This perspective on selection fairness tells a very different story, as the NAS committee noted, because all the disparities in error rates now seem to *favor* Blacks. Looking at the first column for each group, twice as many Whites as Blacks succeed on the job after failing the test (50% vs. 25% are false negatives). Of the individuals who pass the test (the second column), 13% of the Whites but fully 38% of the Blacks do poorly on the job (are false positives).

There is no real technical rationale for the NAS committee preferring the first perspective over the second, and in fact, the committee offered none. Instead, it simply hitched its selective attention to error to a morally charged but unsupported empirical claim (namely, that current burdens on minorities are due to continued discrimination) when it simply stated that "at the very least . . . the inadequacies of [selection] technology should not fall more heavily on the social groups already burdened by the effects of past and present discrimination" (Hartigan & Wigdor, 1989, p. 260).

More importantly, however, it is not clear why anyone would choose to compare these two perspectives on fairness in the first place, as both represent a muddled combination of equally flawed models of selection fairness. Not only do all four constituent models collapse meaningful differences in performance on the predictor and criterion, making lower scoring groups appear more similar to higher scoring groups than they really are, but they are also mutually contradictory.

The committee's performance-based perspective actually consists of two models. As the notes to Table 2 indicate, the first row represents the conditional probability model. That model advocates equal rates of selection from among good workers, TP/(FN + TP). The second row of the table represents the converse probability model, which advocates equal rates of rejection among poor workers, TN/(TN + FP). The rates in Table 2 indicate that both of these standards of fairness have been violated, as the committee's analysis suggests. The committee erred, however, in leaving the impression that both models of fairness can be satisfied at the same time. The literature on test fairness showed long ago that they cannot be satisfied simultaneously at the same cutoff scores (except in certain improbable circumstances). This means that fairness to good Black workers ensures unfairness to poor Black workers, and vice versa.

The committee's prediction-based perspective also represents two mutually contradictory models of fairness. As the notes to Table 3 indicate, the first column for both Whites and Blacks represents the converse equal probability model, and the second column represents the equal probability model. The former seeks equal rates of job failure for Black versus White rejectees, TN/(TN + FN). The latter seeks equal rates of job success for Black versus White hires, TP/(TP + FP). Because they cannot be satisfied simultaneously, fairness to Black acceptees requires unfairness to rejectees, and vice versa.

It is no wonder that the committee gave no technical rationale for favoring the performance-fair over the prediction-fair perspective. The various error rates it invoked provide none, because they all contradict each other. What the NAS committee actually did, in effect, was to marshall the various rates in an ad hoc way to rationalize eliminating disparate impact through racial preferences. Its analysis of the two perspectives actually represents an implicit model of fairness that assigns a higher social value (subjective utility) to avoiding false negatives than false positives and a lower value to avoiding Black false positives than White ones. Such value judgments about whose interests ought to be favored can, and should, be made explicit, for example, by using the expected utility model

of selection fairness that Jensen (1980, p. 409) reviewed. Its statistical logic and methodology for maximizing the overall utility of a selection process require that subjective utilities be specified and quantified for all four quadrants of hits and misses, in this case for both Whites and Blacks.

Just as the committee's rationale for performance-based score adjustments illustrates its selective attention to error, so, too, does its analysis of their effects. The committee concluded (Hartigan & Wigdor, 1989, p. 264) that the expected declines in aggregate worker performance due to performance-fair adjustments would be small and were clearly justified by the dramatic increases in minority referral rates they produce. A different, more sobering view of the social value of increasing prediction error in the service of performance-based fairness is provided by a look at the changing balance of false positives to false negatives, something the committee failed to report.

Raw data are necessary for calculating performance-fair score adjustments, but the effects of those adjustments on error rates can be well simulated using the frequencies in Table 1. The results are presented in the last two columns of Tables 1–3. They satisfy the conditional probability model of selection fairness, which requires equal rates of false negatives among so-called good Black and White workers. In this example, Whites' scores remain unchanged.

Looking at Table 1 (first row), the conditional probability model would be satisfied if score adjustments shifted six of the eight Black false negatives into the true positive category. This would almost double the number of good Black workers who pass the test (from 8 to 14) and cut by three fourths the number of Black false negatives (from 8 to 2).

However, the same score adjustments that reduce the rate of false negatives also increase the rate of false positives among Blacks, because they bump many true negatives into the false positive category (see second row). About one half (12) of the 24 Black true negatives would become false positives (an increase from 5 to 17) after score adjustments. There would thus be at least two new Black false positives (Blacks who pass the test but perform poorly on the job) for every Black false negative eliminated. In this example, then, Black true positives would almost double in number, but Black false positives would more than triple. The number of Blacks correctly classified would fall from 32 (71%) to 26 (58%).

Table 2 shows how adjustments change the two error rates for Blacks. Although the rate of false negatives falls from 50% to 13%, the rate of false positives increases from 17% to 59%. The committee stressed the benefits of the former but largely ignored the costs of the latter.

If an employer hired everyone in Table 1 who passed the test after score adjustments (but no one who failed), false positives would assume special importance. About two thirds of all poor workers (17 out of 26) would now be Black, despite Blacks composing less than one third of all workers hired (31 of the 100). Turning to Table 3, we see that relatively few of the White hires (13%) would fail on the job (be false positives), but more of the Black hires would fail (55%) than not. Being Black would now be strongly associated with being a poor worker.

The committee's attention to the consequences of score adjustments was highly selective in yet another way. It spoke repeatedly of a need to reconcile the social goals of economic productivity, on the one hand, and minority opportunities on the other, and it analyzed the decrements in predictive validity that accompany increases in minority hiring under different score adjustment scenarios (selection ratio, percentage minority applicants, validities, etc.). However, the committee implicitly gave zero weight to individual rights and the interests of higher scoring groups by not including them in any analyses.

As Blits and Gottfredson (1990b) described, the committee discussed such rights and interests mostly to minimize their import and legitimacy in American political and social life. However, for every Black or Hispanic referred or selected due to race-conscious scoring, there is a more highly skilled White or Asian who is not—and perhaps many more who are also passed over in the queue to reach that one minority individual. Scientific balance requires that these legitimate social values be accorded more serious attention in cost–benefit analyses of employment policy.

The New Cycle of Selective Science

The Civil Rights Act of 1991 outlawed race-norming, but it did nothing to relieve the pressures that led to it in the first place. If anything, the Act intensified them by increasing plaintiffs' incentives for bringing lawsuits and employers' costs of losing them (Gottfredson & Blits, 1992). Hence, the ill-fated history of race-norming may be only prologue to the future.

The adverse-impact problem is once again a multiple-symposia, standing-room-only topic at national meetings of the Society for Industrial and Organizational Psychology (Division 14 of the American Psychological Association), and test score *banding* has become the latest hope for a technical solution to that problem (e.g., Cascio, Outtz, Zedeck, & Goldstein, 1991; Sackett & Roth, 1991; Schmidt, 1991). Like race-norming, however, banding provides but a technical pretext to equalize hiring rates by race.

Banding's ostensible technical purpose is to avoid giving undue weight to small differences in test scores, because, we are told, attention to unimportant differences unnecessarily restricts the employment opportunities of lower scoring groups. The proffered solution to tests' less-than-perfect reliability is to band or group together as equal (essentially, to rescore as identical) all "individual scores that are not statistically reliably different from [the highest one]" (Cascio et al., 1991, p. 236).

There are four major forms of banding, depending on how the bands are used (fixed or sliding) and how individuals are selected from them (randomly or minority preference). Only the sliding-band, minority-preference form of banding substantially eliminates disparate impact, and, not surprisingly, it is the form most often advocated.

That form of banding selects minority applicants before others in the band (all of whom may now be treated as equally able as the highest scorer in the band), and then slides the band farther down the continuum of scores each time the top scorer is selected (and thus removed from the band), allowing the band to incorporate a new layer of scores at its lower extreme. Sliding bands provide a way of adding (lower scoring) Blacks and Hispanics to a band (a minority preference may have already depleted the band of minorities) without having first to empty it of (higher scoring) Whites or Asians. The wider the band, the greater the opportunities to equalize hiring rates by race.

Bandwidth is determined by the reliability of the test and the statistical significance level one sets, larger standard errors of difference and higher levels of significance both producing wider bands (Murphy, 1994). For example, the bandwidth for the Wonderlic Personnel Test (reliability .88) would be 0.96 *SD* if one used a 95% confidence interval, but 0.81 *SD* for a 90% interval and 1.26 *SD* for a 99% interval. A less reliable test would have wider bands and a more reliable one narrower bands. It is not uncommon for bandwidths to be as wide as even the largest mean racial differences on a test (as is the case for the Wonderlic with its mean Black–White difference of 0.97 *SD*), thus illuminating how banding can produce racial parity in selection despite large skill gaps by race.

Bandwidths of around 1.0 *SD* are troubling, however, because they equate such a wide range of skill levels. They mean, for example, that individuals at the 50th percentile on a test could be grouped together with and be treated as equivalent to individuals scoring at the 84th percentile, even though the latter often perform substantially better in training and on the job (e.g., Schmidt & Hunter, 1981). Collapse of such substantial skill differences leads one to suspect a fundamental flaw in the procedure's rationale.

In fact, banding rests on a radical reconceptualization of how to judge whether a test score difference (say, of 0.5 *SD*) is meaningful and therefore should be taken seriously in selection. Personnel psychology traditionally has focused on the impact that such a difference in scores can be expected to have on job performance (utility being judged from the test's predictive validity). Banding ignores job performance altogether and advances a very different proposition, namely, that a particular difference in test scores is meaningful only if we can be convinced that two applicants whose test scores differ by that amount actually differ on the trait being measured (the answer depending on the test's reliability and how certain we choose to be).

In technical terms, banding requires that we treat as equivalent any two individuals whose observed scores are not statistically significantly different. In practical terms, it requires that we always assume that two applicants who have different test scores actually have equal skills (have the same true scores), unless the difference is so big that it would occur by chance no more than 5% of the time (if we select a .05 significance level) if our assumption of equal skills is indeed true—which, of course, it may not be.

We know, however, that the higher the test score, the higher job performance tends to be. The applicant with the higher observed score always has a higher probability (not certainty) of performing well on the job, and employers maximize productivity gains (which can be considerable) by always betting on the higher score (e.g., Schmidt & Hunter, 1981). Not surprisingly, score differences typically must be much, much larger under banding than under utility models to qualify as meaningful, which allows banding advocates to dismiss as nonsignificant many of the skill differences they themselves concede have practical import in the workplace.

What banding implicitly requires, then, is that we consistently favor one type of decision error over another. Namely, if we must err in deciding whether two applicants differ in true scores, we must err on the side of wrongly concluding the two are equally skilled rather than falsely judging one superior to the other. In a top-down selection system, such a decision bias obviously favors less skilled individuals and groups, who are already advantaged by less-than-perfect test reliability and validity. (As one wag has put it, "An unreliable test is the best affirmative action program.")

In short, banding requires that employers knowingly refrain from favoring the applicants with the higher odds of success on the job, except when the odds for two applicants are especially divergent—a selection strategy guaranteed to depress expected levels of job performance. From the applicants' point of view, banding is a form of handicapping that hobbles the more skilled.

Although race-norming and banding differ, their ostensibly scientific rationales are thus disturbingly similar. In both cases, proponents invoke the imperfections of tests, not to encourage better measurement or more accurate decisions (say, by improving reliability or adding more predictors to more fully tap the criterion space), but rather to justify *increasing* errors of prediction and classification in order to advantage lower scoring groups. In both cases, technical expertise is turned to disguising a serious social problem (large skill gaps by race) as a technical one and a particular political solution (covert racial preferences) as a scientific requisite. In both cases, proponents wear blinders to the destructive side effects of the practices they advocate.

The Road Ahead

Current group disparities in life circumstances and outcomes are distressing and continue to unsettle our claims to being a fair and decent society. They are also grist for the mill of political conflict and social unrest.

Personnel-selection psychology has learned in the past two decades that imperfect tests are not the source of subgroup differences in tested skills and abilities and also that more accurate selection is no solution (Schmidt, 1988). Yet the field continues to act as if it were by continuing its dogged search for a strictly measurement solution to disparate impact. There will be none, however, because racial differences in test scores arise principally from racial differences in job-related skills and abilities.

Papering over these serious skill gaps with score adjustments in order to satisfy an impossible legal mandate merely impedes the search for possible solutions elsewhere (say, training to raise skills or changing law and regulation to rescind that impossible mandate).

Personnel psychology has an important role to play, albeit a more limited one, in reducing disparate impact in employment selection. Although the field cannot eliminate the inherent cognitive demands of jobs, it can identify the less cognitive elements of job performance. By adding less cognitive predictors (e.g., certain personality traits) to current test batteries, selection psychologists can reduce adverse impact while increasing predictive validity for some jobs. The effects will be marginal, however, for the most cognitively complex (and generally most desirable) occupations.

Personnel-selection psychology can also perform an important service by analyzing the full panoply of costs and benefits of different strategies for reducing disparate impact. But the biggest contribution personnel psychologists can make in the long run may be to insist collectively and candidly that their measurement tools are neither the cause of nor the cure for racial differences in job skills and consequent inequalities in employment.

REFERENCES

Blits, J. H., & Gottfredson, L. S. (1990a). Employment testing and job performance. *The Public Interest, 98,* 18–25.

Blits, J. H., & Gottfredson, L. S. (1990b). Equality or lasting inequality? *Society, 27,* 4–11.

Cascio, W. F., Outtz, J., Zedeck, S., & Goldstein, I. L. (1991). Statistical implications of six methods of test score use in personnel selection. *Human Performance, 4,* 233–264.

Civil Rights Act of 1964, Pub. L. No. 88-352, 78 Stat. 243 (1964).

Civil Rights Act of 1991, Pub. L. No. 102-166, 105 Stat. 1071 (Nov. 21, 1991).

Gottfredson, L. S. (1988). Reconsidering fairness: A matter of social and ethical priorities. *Journal of Vocational Behavior, 33,* 293–319.

Gottfredson, L. S. (1990, December 6). When job-testing "fairness" is nothing but a quota. *Wall Street Journal,* p. A18.

Gottfredson, L. S., & Blits, J. H. (1992). Legislated lawlessness on civil rights. *Delaware Lawyer, 10*(2), 20–23.

Griggs v. Duke Power Co., 401 U.S. 424 (1971).

Hartigan, J. A., & Wigdor, A. K. (Eds.). (1989). *Fairness in employment testing: Validity generalization, minority issues, and the General Aptitude Test Battery.* Washington, DC: National Academy Press.

Humphreys, L. (1989, July 7). Fairness in employment testing [Letter to the editor]. *Science,* p. 14.

Hunter, J. E. (1983). *Overview of validity generalization for the U.S. Employment Service.* Washington, DC: U.S. Department of Labor, Employment and Training Administration, Division of Counseling and Test Development.

Hunter, J. E., & Schmidt, F. L. (1976). A critical analysis of the statistical and ethical implications of various definitions of "test bias." *Psychological Bulletin, 83,* 1053–1071.

Jensen, A. R. (1980). *Bias in mental testing.* New York: Free Press.

Murphy, K. R. (1994). *Potential effects of banding as a function of test reliability.* Manuscript submitted for publication.

Pearlman, K. (1994). *Job families in the United States Employment Service: Review, analysis, and recommendations* (Report prepared for the United States Employment Service, Contract No. 92-442). Washington, DC: U.S. Department of Labor.

Petersen, N. S., & Novick, M. R. (1976). An evaluation of some models for culture-fair selection. *Journal of Educational Measurement, 13,* 3–29.

Sackett, P. R., & Ostgaard, D. J. (1994). Job specific applicant pools and national norms for cognitive ability tests: Implications for range restriction corrections in validation research. *Journal of Applied Psychology, 79,* 680–684.

Sackett, P. R., & Roth, L. (1991). A Monte Carlo examination of banding and rank order methods of test score use in personnel selection. *Human Performance, 4,* 279–295.

Schmidt, F. L. (1988). The problem of group differences in ability test scores in employment selection. *Journal of Vocational Behavior, 33,* 272–292.

Schmidt, F. L. (1990, April). *SIOP Symposium: Affirmative action in the 1990's.* Paper presented at the fifth annual conference of the Society for Industrial and Organizational Psychology, Miami Beach, FL.

Schmidt, F. L. (1991). Why all banding procedures in personnel selection are logically flawed. *Human Performance, 4,* 265–277.

Schmidt, F. L., & Hunter, J. E. (1981). Employment testing: Old theories and new research findings. *American Psychologist, 36,* 1128–1137.

Schmidt, F. L., Ones, D. S., & Hunter, J. E. (1992). Personnel selection. *Annual Review of Psychology, 43,* 627–670.

Sharf, J. C. (1988). Litigating personnel measurement policy. *Journal of Vocational Behavior, 33,* 235–271.

Sherwood, O. P. (1990). Court innocence. *Society, 27*(3), 24–25.

Tenopyr, M. L. (1990). Fairness in employment testing. *Society, 27*(3), 17–20.

Wigdor, A. K., & Garner, W. R. (Eds.). (1982). *Ability testing: Uses, consequences, and controversies: Part 1. Report of the committee.* Washington, DC: National Academy Press.

Wigdor, A. K., & Hartigan, J. A. (1990). The case for fairness. *Society, 27*(3), 12–16.

Wigdor, A. K., & Sackett, P. R. (1993). Employment testing and public policy: The case of the General Aptitude Test Battery. In H. Schuler, J. L. Farr, & M. Smith (Eds.), *Personnel selection and assessment: Individual and organizational perspectives* (pp. 183–204). Hillsdale, NJ: Erlbaum.

Correction to "Summary Report of Journal Operations, 1993"

In the "Summary Report of Journal Operations, 1993" (*American Psychologist,* 1994, Vol. 49, No. 7, pp. 669–670), the numbers given for *Health Psychology* under "Manuscripts" were based on new manuscripts submitted in 1993 only, not on all editorial activity during 1993, which would include editorial decisions on manuscripts submitted prior to 1993. The numbers for 1993 editorial activity should be as follows: No. received = 236, No. accepted = 41, No. pending = 52, and % rejected = 80.