Journal of Vocational Behavior 33, 424–451 (1988)

COMMENTARY

Can We Count on Muddling through the *g* Crisis in Employment?

ROBERT A. GORDON

Department of Sociology, Johns Hopkins University

MARY A. LEWIS

PPG Industries, Inc., Pittsburgh, Pennsylvania

AND

ANN M. QUIGLEY

City of Tulsa, Oklahoma

The fragmentation of academic disciplines handicaps efforts, educational and otherwise, to deal rationally with problems arising from group differences in general intelligence. In personnel psychology, eyes tend to be on the courtroom, but the classroom may prove the more telling arena. Perhaps equally serious is the failure of each discipline to reckon with strains outside of its immediate province in calculating how much latitude exists for errors of its own that might add to those strains within the polity. One such error would involve doing away with selection tests, thereby compromising unwittingly the fundamental principle of merit. Open discussions like those in this special issue are essential if such blunders are to be avoided. This article illustrates these points through comments on the moral, scientific, and legal concerns addressed by the contributors, and especially through a critique of Seymour's (1988) new analyses, which purport to reveal unsuspected racial unfairness in tests. © 1988 Academic Press, Inc.

A sociologist's view. Subgroup differences in *g*, the *g*eneral intelligence factor, confront our nation with problems for which it is neither ideologically

Please address correspondence concerning this article to Robert A. Gordon, Department of Sociology, Johns Hopkins University, Baltimore, MD 21218. Please address correspondence concerning Lewis (1988) to Mary A. Lewis, Director, Assessment & Organization Development, PPG Industries, Inc., One PPG Place, Pittsburgh, PA 15272. Please address correspondence concerning Quigley (1988) to Ann M. Quigley, Testing Coordinator for the City of Tulsa, Personnel Department, 200 Civic Center, Tulsa, OK 74103.

0001-8791/88 $3.00

nor scientifically well prepared, creating a crisis. There is no single discipline specializing in these problems; consequently, relevant expertise, although often of a high order, is fragmented by disciplinary boundaries and thus rendered less effective for maintaining scientific coherence and for educating future elites. Sociology, normally hospitable to the study of broad social problems, is disabled for this vital educational task by the long-standing hostility to individual differences, to IQ measurements in particular, and to capitalism that many eminent persons associated with the discipline have remarked upon (Homans, 1984, p. 347; Horowitz, 1987; Lipset, 1981; Lipset & Ladd, 1972; Simpson, 1980, p. 287; Watkins, 1982). To many sociologists, capitalism is what private sector industrial psychologists too willingly serve, as though economic efficiency were not also a serious concern in Communist China and the USSR.

Problems involving racial groups in our society are typically attributed in sociology entirely to capitalist society, to the "institutional racism" embedded in that society's structure (e.g., Beeghley & Butler, 1974; Mercer, 1973), and to the faulty attitudes of persons reared in the contexts of the first two. The relevance of disconfirming empirical evidence is soon lost as sociologists, and those whom they have educated, climb away from it up this ladder of abstraction toward social structure. Despite the failure of sociology to pursue its broad scientific mandate, sociological explanations concerning racial matters receive more attention in policy discussions than they deserve, perhaps only because they offer the most easily understood formulations and the least touchy rhetoric.

Militating also against a more unified education and better preparation of future leaders concerning what may well prove to be the nation's most serious long-run problem is an aversion to the emotional overload that might overwhelm and perhaps demoralize members of any single discipline that chose to accept a more comprehensive intellectual responsibility for that problem in its various guises. Indeed, efforts to breach disciplinary boundaries in these matters are apt to be greeted as especially impolitic and hence unprofessional, perhaps even as oversteppings of competence. So, most concerned academics toe the line, resigned to the existing, educationally dysfunctional, division of labor.

Besides hindering a much-needed educational initiative, another disadvantage of the balkanization of expertise is that even tough-minded individuals tend to relax their guard in accepting optimistic reports from other disciplines; such reports encourage the feeling that if racial problems and controversies are being handled reasonably well in one's own field, time is on our side. The field of education is the most common source of hope and hence also of unfounded optimism, as Spitz's (1986) review of interventions for raising the intelligence of the retarded—many once ballyhooed in the media—would lead one to suspect.

This special journal issue and its predecessor (Gottfredson, 1986a), and the conferences from which the journal issues have emerged, represent important efforts by aware professionals to transcend disciplinary boundaries, bringing together for wider consumption reports from workers in the vanguard of several fields concerned with racial fairness in employment. The participants include some who are adversaries in the public arena and who accordingly deserve our respect for their willingness to juxtapose their views for common scrutiny. Persons from fields that are still somewhat insulated from the full, societal implications of the racial fairness problem, perhaps because they deal mainly with individuals, as in counseling and cognitive psychology, may question the need for addressing these issues in a challenging way. Let them ponder lessons from other crises, therefore, and, after informing themselves about the contents of these journals, let them ask whether we can count on muddling through this one by permitting just the one side that is critical of testing to prosecute vigorously its case.

Too much is at stake to depend on muddling. Viewed from a sociological perspective on individual differences, the history of racial segregation in the United States can be understood as an attempt to solve the dimly understood problems posed by a large group difference in *g* entirely at the expense of the lower-scoring group (e.g., see the moving story told about his successful father by Allen, 1988). That effort proved morally intolerable, for obvious reasons. There is cause for national pride in evidence that qualified blacks now participate fully in the economy as measured against expectations based on the distribution of *g* (e.g., Gottfredson, 1986b, Table 2). Those fully participating blacks, honest "working stiffs" and DuBois's (1903) "talented tenth," have a special stake in the future of race relations in this country. But stubborn problems posed by the group mean difference in *g* remain, and there is a growing temptation now to solve them entirely at the expense of the higher-scoring group—as though two wrongs would make a right—and thus at the expense of the same meritocratic principle that was indefensibly compromised by forced segregation. This did not work before and it will not work now.

A survivor of constant skirmishing behind scenes, but always upheld under publicity's glare, the meritocratic principle is one of our most cherished norms; whether it is violated de jure or de facto may matter less than that it is violated at all. Even Seymour (1988), a severe critic of employment testing, states that he cares "deeply about merit" (p. 334). Plainly, if respect for merit represents a faulty attitude, it is one widely shared.

Bad science is unlikely to make good law, and so if test critics prevail in the courts and the legislatures, but are wrong about tests, merit will have been severely compromised unthinkingly and unforeseen strains will have been introduced unwittingly into our national life. From a sociological perspective, the overlooked moral in need of emphasis is that the latitude for miscalculation in any one discipline, such as personnel psychology, may be much narrower than its members realize. Stresses from problems attributable to subgroup differences in *g* may be the separate provinces of different disciplines in academia, but they all come home to roost in the same body politic. No potential increment to such stresses can safely be contemplated in isolation, therefore.

Besides violating a major norm on a grand scale and thus straining race relations in ways that can interact unexpectedly with like stresses from other sources, a major policy blunder with respect to employment testing runs the risk of wounding the goose that lays the golden eggs, that is, the American economy (e.g., Hunter & Schmidt, 1982; Schmidt & Hunter, 1980). This serious risk is perhaps the most insidious of all, because the harm occurs too gradually to trigger widespread alarm.

Organization of specific comments. Not every contribution in this issue is equally in need of comment from any one source, and a comment's length does not necessarily index the value of the contribution addressed. As it happens, the largest share of the space available is devoted to Seymour's (1988) article, which contains the newest and least-reviewed data, and so his is discussed last.

ALLEN: MANY IRONIES IN THE FIRE

Allen (1988) draws on ironies from college athletics to counter two misguided policies concerning tests. One consists in failing to use tests where they ought to be used, the other of using tests for the wrong purpose and hence where they ought not be used. The second can be viewed as a nice example of what Gottfredson (1988) refers to as changing the performance criterion—only here the change is in the opposite direction for athletes, to the more *g*-loaded criterion of academic performance from the less *g*-loaded one of athletic performance.

It is a peculiarity of recruitment into certain professional sports that talent is first sifted through college admissions offices. This may have something to do with the founding of professional leagues in those sports at times in which, lacking an infrastructure of minor leagues, they recognized that college athletic programs could serve as ready-made farm systems. In turn, college coaches saw that winning seasons built their reputations and enabled them to recruit talented players seeking the best preparation and showcasing for future professional careers. When enough such vested interests combine, they give rise to a viable social system, however dishonest, which endures providing that its pathologies and contradictions (as Marxists would say) do not become too grossly apparent. Just such a fundamentally flawed, but temporarily flourishing, system is presently taking root in our society, in the form of implicit racial quotas in education and employment.

Although it would make more sense to peg the minimum admissions test scores for athletes to, say, the lower quartile of the college for which they played, this would give academically less demanding colleges the advantage of recruiting from a wider pool of athletic talent than was available to the more demanding ones. Hence, somewhat irrationally, a rather low admissions test cutoff has been set for everyone by Proposition 48, but one which may still impact adversely on some black athletes at some traditionally black colleges, as Allen notes. Thus, efforts to maximize mutually interfering goals with respect to college athletics produce strains, inequities, hypocrisies, and high-minded rationalizations that are often not unlike those spawned by the double standards of affirmative action quotas. The latter carry the more serious threat, however, of stigmatizing all blacks permanently, not just persons temporarily in the athlete's role.

BOLICK: MAKING SENSE WHILE MAKING LAW

Bolick (1988) provides an interpretation of the law that would represent a principled alternative to the single-minded, outcomes-above-all approach by now standard on behalf of plaintiffs in civil rights litigation. As sound as Bolick's approach may be from a constitutional standpoint, it seems to depend for its long-run viability, in a democracy with changing demographics, on one of two possibilities: either (a) finding an acceptable remedy for the problem of group differences in *g*, or (b) promoting a deeper and broader understanding of the reasons why adverse impact to a lower-scoring group must be tolerated for the sake of preserving the merit principle. The first is not by any means visible on the scientific horizon, but the second *is* conceivably achievable by educational means if only the necessity for doing so were clearly perceived. Obviously, these are not comforting thoughts, but that is no justification for not thinking at all. The quotations from blacks included in Bolick's article demonstrate that concerns about protecting merit are shared by members of both races.

GOTTFREDSON: PRIORITIES FIRST

In a thoughtful article relating to test fairness, Messick emphasized the need for exactly the kind of exercise this special issue and Gottfredson's (1988) contribution in particular represent. He stated, citing Churchman (1961) and others in support:

> although consensus is the decision rule of traditional science, conflict is the decision rule of ethics. Since the one thing we universally disagree about is "what ought to be," any scientific approach to ethics should allow for conflict and debate, as should any attempt to assess the ethical implications of science. "Thus, in order to derive the 'ethical' implications of any technical or scientific model, we *explicitly* incorporate a dialectical mode of examining (or testing) models" (Mitroff & Sagasti, 1973, p. 133). (Messick, 1980, p. 1022)

Despite the well-established need for dialectics, there is sometimes an artificial shortage of dialecticians on one side of a crucial issue. For instance, another sociologist, Beer (1987), has remarked:

> In the debate over the effects of reverse discrimination, preferential hiring, and quotas, one surprising fact emerges: social scientists have been almost entirely mute. Twenty years after the enactment of the Civil Rights Act of 1964 . . . there has been no systematic inquiry into the effects of affirmative action on American society, neither its costs to the nation's economy nor its impact on our country's morale. In an age of program evaluation, when most other social experiments are studied almost to death, our profession has shown a resolute ignorance about an extraordinarily controversial policy that has been in place for over two decades. It is as if affirmative action has assumed the status of a religious article of faith, and professionals choose to avoid studying its effects for fear of what they might find. (p. 63)

If Messick is correct, what Beer describes is an unhealthy state of affairs, even allowing for the possibility that he was unaware of relevant work in industrial psychology. Reports of the effects of preferential hiring tend to be few and fugitive, such as one of a 10% reduction in profits following the hiring of 10% minority at every level of an organization (Alderfer, 1982). This was accompanied by the disturbing, but entirely plausible, assertion that "people never tell the truth about race in organizations" (p. 145).

Therefore, Seymour's (1988) claim that there has been substantial progress in job opportunities for minorities and women as a result of çivil rights litigation, particularly where tests have been discontinued, must be viewed accordingly. In the absence of objective evaluations, such claims for progress, even in lower level jobs, can be based only on the numbers of persons hired, not on their work performance. Their credibility may rest on the impression many professionals have that lower level jobs do not depend on *g*, based simply on the fact that the jobs look easy to them. But items on a children's IQ test may also appear so easy to an older person as to seem to require no intelligence at all, although the items happen to be highly *g*-loaded for persons at the appropriate level of difficulty. Thus, the entire topic must be opened up to scrutiny, just as Gottfredson and other contributors to this special issue are doing. "In the areas of its expertise the scientific community has the authority, and the obligation, to help the public to discriminate between rational and irrational views" (Davis, 1986, p. 246).

SCHMIDT: NEVER UNDERESTIMATE A SMALL *r*

Schmidt (1988) has provided an up-to-date summary of the status of his important work on validity generalization with Hunter and other collaborators, including their responses to some recent criticisms based on attempts to revive the hypothesis of a common, spurious bias in predictors and job performance measures. His careful attention to alternative hypotheses in this article is commendable.

Caution against raising false hopes concerning test-score gains must be urged, lest the seeming light at the end of the tunnel cost lead time in confronting the *g* crisis by turning out to be yet another *ignis fatuus* (Spitz, 1986). For example, although slight gains in Scholastic Aptitude Test scores of blacks have indeed been reported, as Schmidt notes, Wainer (1987) demonstrates that, quite aside from the serious problem of differences from year to year in self-selection of the examinees, the key assumption that the numerous nonrespondents to the item from which race is identified have the same test scores as respondents of their own race is untenable. The resulting ambiguity is great enough to dwarf the score changes attributed to blacks in some years, and large enough to equal the total changes observed for blacks over a 6-year period.

Because Seymour (1988) questions the importance of correlations by asking who would depend on stock price predictions that account for "only 4% or 9% or 16% of the variability," it may help to clarify Schmidt's correct interpretation of *r* with a simple example. Imagine trying to predict the toss of a fair coin by tossing another fair coin first. The expected fourfold table will have 50%–50% marginals and will exhibit complete independence between the two variables and a zero correlation. It is worth noting that even with a zero correlation, the expected gains and losses from fair bets are zero, and so one would not be worse off for betting over the long run.

Now suppose that a magic coin were substituted that, although not infallible, produces a .30 correlation with the second coin tossed, thus accounting for 9% of the variability. No matter which direction is chosen for calculating the percentages in this symmetrical table, the True Heads and True Tails cells each contain 65% of their row and column totals and the other two cells each contain 35%. Each percentage difference equals 30% (Davis, 1971, pp. 71–72), and the odds of winning, calculated as the ratio of one of the True cells to one of the False cells (Reynolds, 1977, pp. 34–35), have gone from 1:1 to now 1.86:1. If the payoffs remained unchanged over extended play, one would make a fortune! One would win 65% of the time instead of 50%, a difference of 15%. This difference, when divided by the difference between winning 50% of the time in the perfectly random case of $r = 0$ and 100% of the time when $r = 1.0$, equals .30, which corresponds to both the correlation and the regression coefficient in this symmetrical table. Ergo, Schmidt's important point that a validity correlation of, say, .30 has 30% as much value as perfect validity, is conveniently illustrated for two dichotomous variables. By similar reasoning, it can be demonstrated that at Monte Carlo the house advantage in roulette of 2.7% is equivalent to a correlation of only .027 when a player bets "rouge." Thus are entire industries founded on low correlations. Perhaps now the lawyer's fallacy will take its place beside the gambler's fallacy in textbooks illustrating faulty statistical reasoning.

SHARF: AN INSIDE VIEW FROM INSIDE THE BELTWAY

Sharf (1988) has been well positioned to observe the unfolding of civil rights litigation and its accompanying sets of guidelines for compliance, and has provided a thorough review and analysis of what he views as the corruption of the intent of the original legislation by the adversarial judicial process. Seymour (1988) disputes Sharf's interpretation, holding that thousands of lawsuits have merely secured for women and minorities the protections intended by Congress. The differences between them turn on such issues as what one understands by the phrase "professionally developed ability tests" when such tests produce disparate impact for minority job applicants.

The level of general sophistication concerning emotionalized issues is critical for determining the probabilities that judicial verdicts and legislative decisions will go one way or the other, hence it could well be claimed that employment policies are being determined in the classroom rather than in the courtroom. Note the influence of sociological teachings at several points in the history of the litigation and legislation reviewed by Sharf (1988), particularly in references to such vague concepts as "institutional" and "systemic" discrimination. Inevitable lags in the dissemination of new knowledge mean that the policies of today are being governed by the teachings of yesteryear. James Madison's warning, "A popular Government, without popular information, or the means of acquiring it, is but a Prologue to a Farce or a Tragedy" (Padover, 1953, p. 337), is apropos. The situation that Sharf describes threatens to become such a prologue, demonstrating the need for a more concerted effort in college classrooms to provide leaders of the future with adequate background for addressing civil rights issues.

Without access to quantified research into the effects of preferential hiring, it may never occur to today's educated laity to wonder whether the introduction of an ability test by the employer in *Griggs v. Duke Power Co.* (1971) was solely for the purpose of excluding blacks or simply a response to the more variable applicant pool created by desegregation. Similarly, without experiencing such a situation personally, the laity may not appreciate that whether or not a secretary comprehends the material being typed for an executive may play a major role in how many errors are made, since text that does not make sense would be easily spotted in the former case. Consequently, a more difficult test may, in fact, be job related, although it will seem little related to pushing typewriter keys if the job is misconceived in too narrow terms.

One's attitude toward the above examples is subject to the ambiguities of verbal narrative in ways that would not be the case with hard data.

Whether one sides with Sharf or with Seymour, therefore, will hinge on one's interpretations of similar ambiguities that are inherent in the essentially verbal body of material known as "law." However, Seymour's (1988) article provides an example of his treatment of hard data that can serve as a litmus test of how more ambiguous legal material might be handled by plaintiffs. Reading Seymour's article in conjunction with the critique that follows, therefore, can equip one to decide better between him and Sharf, and also to respond to the question that serves as title for this set of comments.

SEYMOUR: HORIZONTAL v. VERTICAL

Seymour (1988) declares that "skepticism toward the use of tests has now been shown to be justified." He supports this claim with a new study (Seymour, 1989) that he offers as "proof" of the righteousness of his legal crusade and that he uses as springboard for attacking validity generalization through its inability to detect the racial unfairness in tests that he claims to have uncovered. Thus, the scientific linchpin of Seymour's (1988) article is this new way of examining validity studies that furnishes important evidence of "racial unfairness."

Given the importance of the issues, if they prove sound Seymour's new analyses would be a refreshing change from the frequent references to adverse impact with their accompanying innuendoes that, joined with sniping at the methodological assumptions of validity generalization, too often pass for ammunition against testing. Let us, therefore, consider the core of Seymour's recent work carefully, passing over his quibbles with correcting correlations and with assuming linearity (on which see Hawk, 1970; Jensen, 1980, pp. 319–320; Schmidt, Hunter, McKenzie, and Muldrow, 1979; Society for Industrial and Organizational Psychology, Inc. [SIOP], 1987), which supersede his reference to an obsolete claim).

The Source of Data

In brief, Seymour (1989) has reanalyzed data separately by race from Labor Department validity studies of 47 middle to low level occupations. In each study, a composite test score from the Specific Aptitude Test Battery (SATB) found appropriate for the occupation was dichotomized to yield predictions of "poor" and "good" (i.e., less and more satisfactory) workers and trainees, as determined from two sets of supervisor ratings made several weeks apart. Approximately the top two-thirds of ratings in each study were intentionally deemed "good" and the test cut-score was chosen so as to select persons in that category (L. M. Avery, personal communication, September 15, 1988; United States Employment Service, 1970, pp. 50–52). All subsequent analysis derives from the results so classified within two sets of 47 fourfold tables, one set for whites, one set for blacks. Following Seymour, workers and trainees are referred to as "workers."

Seymour's New Study of "Racial Unfairness"

Basic error rates. Seemingly in accordance with standard usage, Seymour (1989) explained the two types of mistakes that a test can make as "the 'false rejection' mistake of excluding 'good workers,' " which he refers to as *underprediction,* and the " 'false acceptance' mistake of selecting 'poor workers' " (p. 21), which he refers to as *overprediction.* He then presented the rates of each kind of error, expressed as percentages, for both races from each of the 47 validity studies. Two kinds of analyses were employed to summarize his statistical findings: (a) mean percentages and significance tests of the differences between them, and (b) what is practically equivalent to racial ratios of each pair of mean percentages (explained below).

Mean percentages. Testing the significance of the average racial difference in the two kinds of error rates, Seymour (1989, Tables 1 and 5, 1988) found them each significant at probabilities less than .000001. As we shall see, these heroic rejections are of null hypotheses no reasonable proponent of testing would ever consider either true or pertinent for judging tests. One assumes that these significance tests are intended for awing lay persons and judges. In *Larry P. v. Riles* (1979), for example, Judge Peckham (1979) seemed overimpressed when he referred in his decision barring IQ assessments of blacks to testimony that the over-enrollments of minority students in special classes for the educable mentally retarded "could not be the result of chance" since there was "less than a one in a million chance" (p. 23) of that occurring under the null hypothesis—as though anyone had ever invoked chance as an explanation of the racial disproportions involved (Gordon, 1980b, p. 212). If judges understood a bit more about statistics, they might be tempted to cite those offering overwhelming evidence for the rejection of meaningless null hypotheses for contempt of court.

Racial ratios. Not resting his case with examining differences, Seymour (1989) presented, in his Tables 3 and 7, the two error rates for each of the 47 jobs for one race expressed as multiples or ratios of the corresponding error rates for the other race, as well as the averages of each set of 47 ratios. The ratios are composed, with respect to which race serves in the denominator, so that a multiple greater than 1.0 implies disadvantage to blacks. Thus, what he called rejection error rates averaged 1.9 times as large for black good workers as for white good workers, and what he called acceptance error rates averaged 1.6 times as large for white poor workers as for black poor workers.

Derivative analyses. Seymour (1988, 1989) then drew upon the data for whites from his Table 5 concerning the percentage of white *poor* workers who pass the test and contrasted it, for each occupation, with the percentage of black *good* workers who pass the test (derived as the complement of the rejection error percentages for blacks in his [1989] Table 1, i.e., as the difference between those rejection error percentages and 100%). Purporting to portray the "extreme nature of the unfairness in these tests" (1989), Seymour combined the data from these sources to report two seemingly sensational findings:

"(a) In 9 of the 47 SATBs, white *poor* workers have a *higher* test passing rate than black *good* workers; and

(b) in 25 of the 47 SATBs, the test passing rates for white *poor* workers are within 10 percentage points of the test passing rates for black *good* workers (p. 39)"

and concluded finally that "it would be difficult to overstate the importance of test unfairness" (1989, p. 46).

Critique and Discussion of Seymour's New Analyses

The established conventions of cross-tabulation. Two-by-two predictive tables such those from the 47 SATB studies yield cells containing False Positive (FP) and True Positive (TP) outcomes in their top row, and True Negative (TN) and False Negative (FN) outcomes in their bottom row. In medical and psychiatric practice the positive condition of interest is usually an unwelcome one, such as disease (Galen & Gambino, 1975) or dangerousness (Gordon, 1977, 1982), but there is no reason not to employ a welcome outcome, such as "good" worker, for the positive condition.

In a classic reference concerning the analysis of cross-classifications, Zeisel (1957) describes "the cause-and-effect rule" for determining the direction in which percentages should be calculated. He states, "it is not a question of which factor *is* the *cause* of the other one, but which factor we wish to *consider* as affecting the percentage distribution which the other factor assumes" (pp. 24–25). Although he also acknowledges that in many tables, "either factor can be usefully considered as the causal one" (p. 25), the examples provided indicate that this option is most available when temporal ordering of the variables is blurred, as is common in sociology, so that no compelling basis for preferring one causal direction over the other exists, aside from what may arise from the purposes of the investigation.

When temporal ordering and causal directionality are salient, investigative purposes are typically organized around and aligned with causal relationships. Hence, the cause-and-effect rule not only governs percentages in those cases, it is also decisive for shaping the analysis. Citing Zeisel's rule, Davis (1971) concisely describes it, and with it the correct practice, thus: "when two [variables] have an asymmetrical causal relationship, one should use the cause (independent variable) as a base and calculate the percentage showing the effect (dependent variable)" (p. 70).

Where prediction is concerned, temporal priority and hence the putative causal role is understood to belong to the predictor, even when the validation study is a concurrent one. Therefore, for the fourfold tables described above percentages should be calculated *horizontally,* so that Positive and Negative predictions each sum to 100%. With this background in mind, let us return to Seymour's study.

Seymour's error rates do not refer to predictions. Seymour's (1989) case depends entirely on his *vertical* method of defining acceptance and rejection errors, which is unorthodox. In his terms, but using conventional nomenclature that he did not spell out clearly, he defined rates of acceptance errors as FP/(FP + TN) and of rejection errors as FN/(FN + TP). Unfortunately for unwary readers, one must forcibly break set in order to realize the difference between his and the usual definitions of the named error rates.

Normally, acceptance and rejection errors are understood to be the False Positive and False Negative proportions, respectively defined as FP/(FP + TP) and as FN/(TN + FN). In medical diagnosis, the complementary traditional proportions (i.e., TP or TN instead of FP or FN in the numerators) represent the *predictive values* of Positive and Negative test results (Galen & Gambino, 1975). Clearly, whatever Seymour's error statistics are, they are not descriptions of predictions. To qualify as descriptions of the lack of success of predictions, the outcomes predicted must be homogeneous in the denominators, that is, either all P or all N. The prediction, after all, is either "P" or "N," just as in tossing a coin the prediction is either "heads" or "tails," not both.

Schreier (1957, p. 128) has contrasted the interpretations appropriate to both Seymour's and the standard methods of calculating percentages when the association between variables is causal that, if reworded to apply clearly to either of Seymour's situations (e.g., here his acceptance error), would read as follows: Seymour's interpretation indicates the chance that job failure was or was not *preceded* by the antecedent test pass; the standard interpretation indicates the proportion of cases in which passing or failing the test were *followed* by job failure or success. The emphasized indications of temporal sequence make clear that references to Seymour's rates as "predictions" are misnomers, and that for describing prediction he has used the wrong conditional probability.

Seymour's erroneous conception of prediction renders his claims concerning under- and overprediction wrong and seriously misleading. In view of the hopelessly defective nature of Seymour's (1989) conception of "prediction," his conclusion that tests in the 47 studies "systematically underpredict black job performance and overpredict white job performance'' (p. 50) is an especially misleading statement. Not only is it well established that black performance is not underpredicted (i.e., not underestimated by tests) in regression studies of test fairness according to the accepted Cleary (1968) definition of fairness (SIOP, 1987; Wigdor & Garner, 1982), but when discrepancies occur they typically involve slight overprediction for blacks (i.e., a bias in their favor).

In regression studies, under- or overprediction for a group refers to the net or average outcome in performance, under- and overpredicted performances being averaged together. In fourfold tables, under- and overprediction refer to the numbers of under- and overpredicted performers, which cannot be averaged together to produce a single number because there is no way to take the magnitude of each misprediction into account. Hence, the two forms of misprediction must be considered separately for such tables.

Under- or overprediction in both the regression sense and the traditional categoric sense of medical diagnosis is evaluated for any individual by comparing that person's actual performance or outcome with his predicted performance or outcome. Consequently, the meaning of under- or overprediction is essentially the same for individuals and groups in both types of analysis. But Seymour's definitions have *no common meaning for individuals and groups*. Although his numerators do have the same meaning for individuals as they would have in medical diagnosis, his group definition entails dividing by denominators that mix Positive and Negative predictions; both predictions are mutually exclusive and cannot hold simultaneously for any individual.

The proper measure of underprediction requires comparing the False Negative (i.e., false rejection) proportions of blacks and whites, as defined above. Measured properly, blacks are underpredicted (have higher rates) relative to whites in only five of the 47 studies and not in 43 of 47 instances as claimed by Seymour (1988, 1989). In the remaining 42 studies, it is the whites who are underpredicted. Similarly, properly measuring overprediction requires comparing the False Positive (i.e., false acceptance) proportions of blacks and whites. This shows that whites are overpredicted (have higher rates) relative to blacks in only six of the 47 studies and not in 41 of 47 instances as claimed by Seymour (1988, 1989). In the remaining 41 studies, it is the blacks who are overpredicted. The few instances of underprediction of blacks or of overprediction of whites never approach statistical significance. (Only one χ^2 even exceeds 1.0.)

Properly analyzed, therefore, results from these studies are entirely consistent with findings from regression studies: blacks are typically not underpredicted, but are typically overpredicted (Hunter, 1983; Lewis, 1988, Table 1; Quigley, 1988, Tables 1 and 2; SIOP, 1987; Wigdor & Garner, 1982). Seymour's rejections of null hypotheses concerning mean differences between blacks and whites in each of his two rates are therefore pointless insofar as the issues of under- and overprediction are concerned.

Sensational derivative analyses refuted. Seymour's (1988, 1989) most sensational claims were (a) that in 9 of 47 SATBs white poor workers had higher passing rates than black good workers, and (b) that in 25 of 47 SATBs the test passing rates for white poor workers were within 10 percentage points of the test passing rates for black good workers. Basing these comparisons on percentages calculated in the conventional way reveals that in no case does the white False Positive rate exceed the black True Positive rate, and the two rates are never within 10 percentage points of each other.

Do Seymour's Rates Have Any Meaning at All as Measures of Racial Unfairness?

What then are we to make of Seymour's (1989) analyses, if they do not accomplish what unwary readers might be led to expect? Have they any meaning as evidence of *racial* differences in the way that tests work (as claimed in his title) that might serve as justification for the litigation he threatens to pursue? Certainly, Seymour's statistics do not qualify as predictions, but what if they are regarded simply as observations? Is it realistic to hope that his analyses will be withdrawn once their defects are noted?

On the contrary, one must anticipate Seymour's capitalizing on the degree of judgment that methodology texts acknowledge is needed when percentages are applied to contingency tables, even as the texts strive to impart wisdom without dogma. In the absence of absolute, inflexible dogma, lawyers can find ambiguity when in fact none exists for the expert (one reason, perhaps, that unlike Bolick, 1988, and Sharf, 1988, Seymour prefers leaving the door to litigation wide open). The preceding questions must be addressed, therefore, in preparation for the day when Seymour brings his misleading evidence to court, his own experts in tow, hoping to find a judge who will decree how contingency tables ought to be percentaged and interpreted. Who knows but what some judge will try for a Solomonic resolution—in actuality a Solomonic fatal compromise—by deciding that the two directions of percentaging must somehow be given equal consideration. (Note: Seymour's percentages represent simply an aspect of adverse impact due to group differences in level of tested ability and so are already taken into account by law. See Appendix.)

Brief summary of analyses in Appendix. Examination of the two correlation matrices for blacks and whites of associations among cell and marginal percentages for the 47 fourfold tables reanalyzed by Seymour (1988, 1989) reveals quite similar behavior of the tables for both races. As passing rates vary, cases flow heavily between True Positive and

True Negative cells in both races, producing the effects on error rates on which Seymour relies.

The ultimate demonstrations that Seymour's (1988, 1989) contrasts between black and white error rates have nothing to do with race, but only with differences in mean levels of tested ability between blacks and whites, are contained (a) in regression analyses that successfully model his effects entirely from variation in passing rates within either race alone, and (b) in an analysis based on imposing the passing rate marginals of each race in turn on the True Positive and True Negative rates of the other, which also successfully models his effects, again in either race alone, entirely from variation between races in passing rates.

Final Remarks on Seymour's New Evidence

Whatever one's attitude toward Seymour's error rates, it should be evident that they lack standing as evidence in civil rights litigation because they have nothing to do with race per se and because, instead of showing that tests work differently for blacks and whites, they represent simply one more instance of the unhappy fact that tests work just the same for both groups (ignoring the fact that blacks benefit slightly in selection decisions from overprediction). All of Seymour's effects were demonstrated to derive entirely from differences in level of ability between blacks and whites and could be expected to occur between two black groups or between two white groups displaying the same difference in test scores and hence passing rates (see Appendix). Therefore, his results are simply a restatement of adverse impact. If there is unfairness in these facts, that unfairness lies in the realities of group differences in ability and not in the tests, just as it might lie in the realities of individual differences. Because the evidence emerges from differences in score and ability and not from race, if Seymour's new evidence were to be accepted in civil rights cases, the question would legitimately arise as to why, say, the middle third of white scorers did not enjoy equal protection against the top third of white scorers on exactly the same basis, and, by the same reasoning, the bottom third against the middle third. This would constitute a rejection of the idea of merit itself and hence the ultimate *reductio ad absurdum*.

Since Seymour's (1989) study turns out to be of the effects of shifting marginals on contingency tables, not of race, it simply represents one more attempt to parlay the unfortunate fact of adverse impact of group differences in ability into a broad indictment of our society, in this case, of the occupational world and the discipline of industrial psychology, just as Mercer (1973) attempted with the educational realm and the disciplines of school and educational psychology (Gordon, 1980a). That such scientifically flawed efforts may proceed from good intentions, and sometimes pay off in court (e.g., *Larry P.*), does not lessen their corrosiveness to public morale or their long-range risks to race relations and major institutions when they succeed in misdirecting public policy.

APPENDIX

What are the systematic properties of Seymour's error rates? The so-called false rejection and false acceptance rates that Seymour (1989, Tables 1 and 5) reports happen to represent the complements (i.e., differences from 100%) of statistics known in medical practice, not as predictions, but as the *sensitivity* and the *specificity* of a diagnostic test (Galen & Gambino, 1975). Sensitivity is defined as TP/(FN + TP) and specificity as TN/(FP + TN). These statistics differ from Seymour's only in the choice of term from their denominators placed into their numerators, and so each would correlate perfectly but negatively with the complementary statistic of Seymour's. In view of the perfect correlation, each complementary pair contains the same information in reverse form, and so whatever holds true for relationships affecting the medical statistics applies also to Seymour's error rates.

Sensitivity and specificity tend to be inversely related to each other, so that increases in one are accompanied by decreases in the other, and these are linked to changes in the True Positive and True Negative cells: "Positivity in disease [i.e., sensitivity] is coupled *inversely* to negativity in health [i.e., specificity]. If there are more true-positive results in diseased subjects, then we are likely to find a smaller number of true-negative results in healthy subjects" (Galen & Gambino, 1975, p. 12); "the higher the sensitivity, the lower the specificity" (p. 50). Applied to Seymour's statistics, these comments suggest the presence of systematic relations of an inverse sort between his statistics that would account for the black–white differences he obtained.

Measurement specialists will recognize in sensitivity and specificity, and hence in Seymour's (1988, 1989) complementary two rates, applications mathematically identical to the key probabilities in what Petersen and Novick (1976) termed the "conditional probability" (p. 10) and "converse conditional probability" (p. 15) models of fair selection. Each model was intended to produce group-specific cut scores that would equate one of Seymour's two kinds of selection errors across all groups. In their discussion, Petersen and Novick called attention to the basic contradiction involved, as Seymour did not, in the fact that equating both errors by manipulating each group's predictor cut score was impossible, because an adjustment in a group's cut score that would reduce one of the errors would increase the other (pp. 16, 23). This is consistent with the more general observations of Galen and Gambino (1975). However, although Petersen and Novick's contradiction exposes essentially the same important property of Seymour's rates, if left just as it stands it does not automatically dispose of those rates since Seymour's intention is not simply to meddle with cut scores but to do away with the tests. Therefore, a more detailed clarification of the meaning and properties of Seymour's rates that explicitly traces their implications beyond Petersen and Novick's contradiction is needed. Such a clarification follows, based on Seymour's own empirical data.

Examining the aggregate contingency tables for whites and blacks for insights into Seymour's statistics. Figure 1 presents cross-tabulations of the aggregate data separately for blacks and whites, from the 47 studies Seymour reanalyzed. Percentages have been calculated in three separate ways: (a) so as to sum to 100% for each *row,* as would be standard practice; (b) so as to sum to 100% for each *column,* as in Seymour's analyses; and (c) so as to sum to 100% after dividing cell frequencies by the grand *total,* which converts all cells in cross-tabulations to a common basis. Method c is not considered especially informative, but it is never "wrong" and can be useful.

Method a, labeled row %, shows that whites are underpredicted more than blacks in the False Negative cells and that blacks are overpredicted more than whites in the False Positive cells, contrary to Seymour's claim, as was to be expected from published evidence.

WHITE SAMPLE DATA		JOB PERFORMANCE		
		Poor (Fail)	Good (Succeed)	Row Total
	Succeed	1,160	3,980	5,140
	Row %	22.6%	77.4%	72.3%
	Column %	54.8%	79.6%	(Pass%)
	Total %	16.3%	55.9%	
TEST		False Positive	True Positive	
PREDICTION	Fail	955	1,019	1,974
	Row %	48.4%	51.6%	27.7%
	Column %	45.2%	20.4%	
	Total %	13.4%	14.3%	
		True Negative	False Negative	
	Column Total	2,115 29.7%	4,999 70.3%	7,114 100%

Whites: Efficiency = 69.4% r_{phi} = .25 r_{tet} = .43
Blacks: Efficiency = 64.7% r_{phi} = .29 r_{tet} = .44

BLACK SAMPLE DATA		JOB PERFORMANCE		
		Poor (Fail)	Good (Succeed)	Row Total
	Succeed	512	1,116	1,628
	Row %	31.4%	68.6%	49.3%
	Column %	33.4%	62.9%	(Pass%)
	Total %	15.5%	33.8%	
TEST		False Positive	True Positive	
PREDICTION	Fail	1,020	657	1,677
	Row %	60.8%	39.2%	50.7%
	Column %	66.6%	37.1%	
	Total %	30.9%	19.9%	
		True Negative	False Negative	
	Column Total	1,532 46.4%	1,773 53.6%	3,305 100%

FIG. 1. Aggregate cross-tabulations of test prediction with job performance for blacks and whites in 47 SATB studies.

The percentage differences are about 10% in each case. Certainly, there is no indication here, with percentages calculated in the standard manner, of the unfair discrimination against minorities that one would have assumed Dunnette (1974) was warning against when Seymour (1989) quoted him to support conclusions drawn from his own peculiar error rates.

Method b, labeled column %, yields white/black ratios of what Seymour called acceptance errors, and black/white ratios of what Seymour called rejection errors that are comparable in magnitude to the averages of such ratios from each of the 47 studies that he examined (Seymour, 1989, Tables 3 and 7). Here, the two ratios are 54.8/33.4 = 1.6 (compared to his 1.6) and 37.1/20.4 = 1.8 (compared to his 1.9). Based on the ratio of means rather than the mean of ratios, and so expressed in similarly aggregated form, Seymour's ratios are 1.5 and 1.8, respectively. Thus, in these two key respects, our aggregation of the data is comparable to Seymour's summary data for the 47 individual studies in his analyses and therefore can be used to examine his method. (For individual jobs, of course, these ratios ranged widely, as ratios are apt to do when their numerators and denominators vary independently. Hence, the results for some jobs appear far more dramatic than those for other jobs, erroneously suggesting that the former would be riper targets for precedent-setting litigation. Because of such variation, the mean of ratios is not necessarily equal to the ratio of the means of its numerators and denominators [e.g., Stanley, 1957]; the latter ratios here also differ from the former as a result of weighting the input from each study according to its sample size rather than equally. This is pointed out in order to make clear the comparability of our analysis based on the aggregated data and the reasons for the slight discrepancies noted above.)

Method c, labeled total %, in which all percentages share the same basis, reveals that Seymour's emphasis on systematic racial differentials in error rates notwithstanding, the percentages of all blacks and of all whites falling in the False cells are rather uniform. The method c percentages in Fig. 1 suggest that the real source of the black–white differences that Seymour considers important stem from between-race differences in the diagonal distribution of percentages in the True cells, which would influence his error rates. But the True cell distributions for these aggregate data are in turn associated with a conspicuous difference of 23 percentage points between blacks and whites in passing (Pass%) and failing (100% − Pass%) the test, that is, with the two right-hand marginal distributions. Contingency tables are well known to be sensitive to shifts in marginal distributions (e.g., Davis, 1971, p. 70; Zelditch, 1959, pp. 138–139).

Indeed, the marginal shifts of the independent variable in Fig. 1 render neither racial sample strictly representative of the combined sample, and each unrepresentative of the other, and so the shifts argue for applying another of Zeisel's (1957) rules. Stated less confusingly by Zelditch (1959, p. 138), that rule is to percentage cells so that they add to 100% at the margin that is arbitrarily unrepresentative. This rule can override the cause-and-effect rule, but in this case the two rules reinforce each other because it is the causal marginal that is "arbitrarily unrepresentative" as a result of sampling two groups that come mainly from different segments of the ability range. The 23 point difference in passing rates signifies a difference of .59 SD between black and white mean scores.

As was expected from the large black–white differences in distribution across True cells in Fig. 1, coupled with the relative lack of black–white differences in the False cells, sensitivity is 16.7 points higher for whites and specificity is 21.4 points higher for blacks. These differences are close to, but somewhat less than, the race difference in right-hand marginal percentages. These are the statistics with which Seymour's are perfectly negatively correlated, and they display the inverse relation in these between-race comparisons that Galen and Gambino (1975) and Petersen and Novick (1976) described. The possibility that the differences of interest to Seymour are driven by the black–white difference in right-hand marginal passing rates needs to be considered further, therefore. Pseudo-racial cross-tabulations for new white samples with the same predictor marginals as were observed for blacks, and for new black samples with the same predictor marginals as were observed for whites, would be one way to proceed. The hypothesis would be that they would yield essentially the same contrasts with the existing samples, and with each other, that Seymour reported. If so, there would be nothing "racial" about the outcomes that figure so prominently in Seymour's arguments.

Lacking pools of whites and blacks from which to select pseudo-racial samples (e.g., Jensen, 1974, 1977), appropriate models for comparison can be formed in two other ways from the 47 studies. One model is based entirely on variation within each race, via simple regression techniques. The other model is based entirely on variation between the two races, and employs the aggregate data by interchanging right-hand marginals while leaving certain internal features of the black and white fourfold tables intact. The resulting analyses are orthogonal to each other, and have the advantage over pseudo-racial samples of reflecting the same associations between test score and job performance that the existing samples exhibit, a property that new samples would not necessarily possess due to sampling variation.

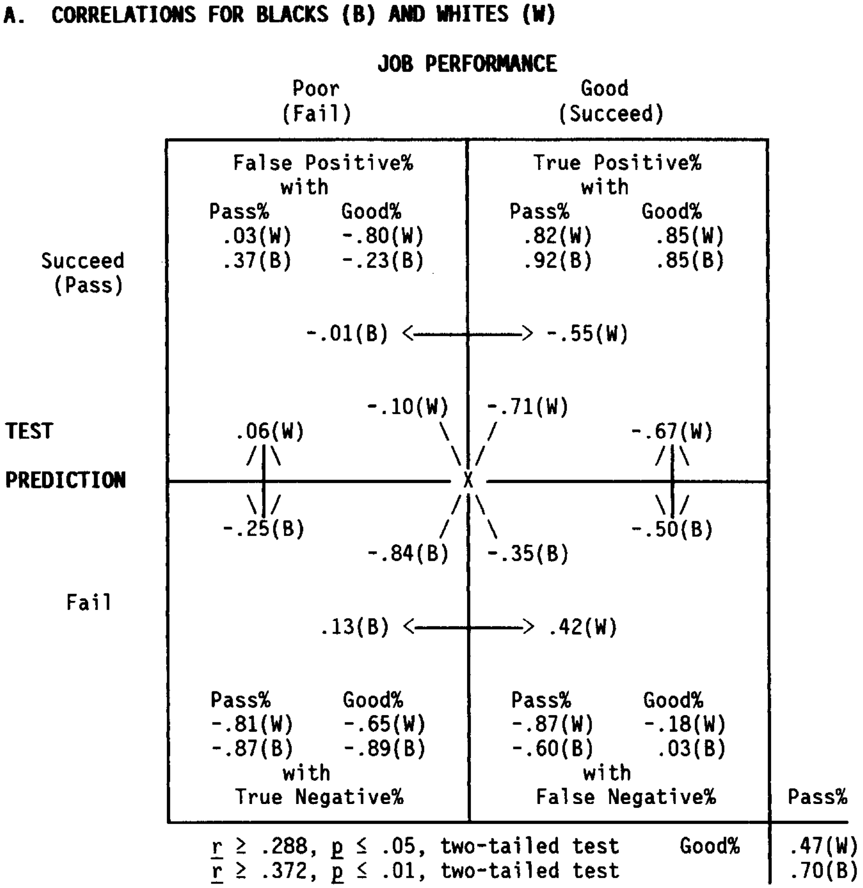
Before leaving Fig. 1, attention should be drawn to the comparability of the three measures of association between predictor and outcome for both races: efficiency, ϕ, and the tetrachoric correlation. Galen and Gambino (1975) define the *efficiency* of a diagnostic test simply as the percentage of all cases falling in the True cells (i.e., "hits"). If the True Positive and True Negative rates of blacks were applied to the white test passing marginals (which would lead to more blacks being exposed to their own higher True Positive than False Positive rate), the modest black–white difference in efficiency would be smaller by 36%; thus, the difference in right-hand marginals accounts for over one-third of the small race difference in efficiencies. Other measures of validity in Fig. 1 are slightly in favor of the black sample. That such highly comparable sets of test data should have become the object of Seymour's ambitious attack suggests that genuinely unfair test results must be in short supply.

Correlational analysis. As background for regression analysis, within-race correlation matrices and necessary summary statistics for blacks and whites are presented, but in somewhat unusual form, in Fig. 2. The unlabeled numbers in Fig. 2A are correlations among cell and marginal percentages calculated to a common basis by method c, above, over 47 fourfold tables for blacks and the corresponding 47 tables for whites.

In Fig. 2A, correlations connected by a line apply to the 47 paired percentages in the two cells whose borders the line crosses. Corresponding correlations for blacks (B) and whites (W) appear at opposite ends of the same line. This presentation aids in mapping the correlations onto the layout of the fourfold table and in comparing black and white correlations for the same pairs of cells. Correlations within cells not linked to lines refer to covariation of the cell percentage with the marginal percentages. Correlations between the two marginal's appear at the lower right corner, outside the four cells.

The within-race correlations of the True Positive and True Negative cells with both marginals, and with each other, are strong and very similar for blacks and whites. Likewise, other black and white correlations in Fig. 2A tend to display the same patterns as to sign (the exceptions occur when both are low and nonsignificantly different from zero) and as to general level of magnitude. Variation between studies around the mean Pass% for whites occurs in a score range far above the test's cut-point, and so tends to have less impact on variation in the Good% than in the case of blacks, whose mean Pass% lies right at the cut-point. Hence, the correlation between the two marginals is lower for whites than for blacks. According to Fig. 2A, contingency tables describing the relation of the test to performance within the 47 jobs work pretty much the same way for both races in their respective score ranges despite consistently large differences between races in right-hand marginal values; this is another indication that the test works the same way for both.

For understanding Seymour's statistics, certain correspondences between black and



B. MEANS AND STANDARD DEVIATIONS OF VARIABLES CORRELATED

	True Positive	False Positive	True Negative	False Negative	Passing Percentage	Good-worker Percentage
White						
Mean%	56.55	15.04	14.09	14.30	71.60	70.86
SD	7.71	4.40	3.52	4.13	6.44	5.80
Black						
Mean%	34.31	15.10	32.01	18.58	49.42	52.89
SD	8.79	3.64	7.64	4.68	9.48	7.60

FIG. 2. Correlations between cell and marginal percentages (total *n* as base) for blacks and whites in 47 SATB studies, and their means and SDs.

white correlations in Fig. 2A are helpful. The percentages of all individuals in any study falling in the True Positive and True Negative cells are powerfully predicted by the percentage scoring above the cut-point (i.e., by Pass%), the first cell positively, the second cell negatively. Consistent with these facts, the two True cells are strongly negatively correlated with each other, much more so than the two False cells. The much greater negative correlation between True cells implies that cases are redistributed so as to remain consistent with the validity of the test as the passing rate varies in response to the ability level of the group tested. There is also a lower but still substantial negative correlation between the True Positive cell and the False Negative cell. The False Positive cell, on the other hand, seems less responsive to changes elsewhere than other cells, which suggests that "qualified" persons who perform poorly do so for nonintellectual reasons. If cell standard deviations are not too different from one another (a problem that regression analysis surmounts), these facts are sufficient to demonstrate that the percentage passing largely determines the relative magnitudes of the specificity (in the left column) and the sensitivity (in the right column) in the 47 studies, and hence also determines Seymour's complementary error rates.

Just as Galen and Gambino (1975) and Petersen and Novick (1976) foretold, and as the between-race picture in Fig. 1 suggests, the empirical relation between specificity and sensitivity within both races is an inverse one, and hence so is the relation between Seymour's two error rates. For whites, his error rates have a correlation of $r = -.11$, and for blacks, whose Pass% has 2.2 times the variance of the whites' Pass% (Fig. 2B), an $r = -.51$. Consequently, when two groups have passing percentages overall (and on the average) that are 3.4 white SD or 2.3 black SD apart, as Fig. 2B indicates, it is reasonable to anticipate that the group with the much higher Pass% will display a much higher rate of what Seymour calls acceptance errors (and whites do, by 21.4 points) and a much lower rate of what Seymour calls rejection errors (which whites also do, by 16.7 points), regardless of whether they are from the same or different races. Such a trade-off is inevitable.

In concrete terms, when validity is at least moderate, as is the case in the 47 SATB studies, the passing percentage, Pass%, acts as the piston in a hydraulic system. As the passing percentage declines, cases flow into the True Negative cell (from the True Positive cell), increasing the denominator (FP + TN) of what Seymour terms the acceptance error rate. According to the correlations in Fig. 2A, the expected value of the False Positive cell is little affected by changes in the passing percentage, and so Seymour's acceptance error rate FP/(FP + TN) systematically declines as Pass% declines, and rises as Pass% increases, mainly as the result of changes in its denominator. Since the passing percentage is 23 points higher for whites than blacks, Seymour's acceptance error rate should be higher for whites than blacks if the test performs the same way for both races. This would yield a white/black ratio of the error rate greater than 1.0, ostensibly favoring the selection of "poor" white workers over "poor" black workers. This is what Seymour found and interpreted mistakenly as evidence that the test works differently for each race.

The correlations in Fig. 2A help trace the effects of changes in the passing percentage, Pass%, on Seymour's rejection error rate in the same manner, assuming for the moment that cell standard deviations are not so different from one another as to becloud the issue. As the passing percentage declines, cases flow out of the True Positive cell and into the True Negative *and* False Negative cells. The numerator, FN, of Seymour's rejection error rate, FN/(FN + TP), therefore rises. Although FN also represents one component of the denominator and is rising, the second component, TP, may be decreasing even more rapidly (note that it is losing cases to two cells, TN as well as FN, and that its positive correlations with a decreasing Pass% are extremely high, especially for blacks). Hence, Seymour's rejection error rate systematically rises as Pass% declines, and declines as Pass% increases (inversely to his other error rate). Once again, since the passing percentage is much higher for whites than for blacks, Seymour's rejection error rate should be higher for blacks than for whites if the test performs identically for both races. This would yield a black/white ratio of the error rates greater than 1.0, ostensibly favoring the rejection of "good" black workers to a greater degree than that of "good" white workers. This, too, is what Seymour found and interpreted mistakenly as evidence that the test works differently for each race.

Note that a less valid selection device would clearly favor the nonselection of good workers of both races, and hence reject proportionately more good white workers than good black workers; this eventuality does not seem to concern Seymour (1988, 1989). His cleverness lies in juxtaposing his two conditional error rates for blacks and whites, which are always linked in a trade-off *within* any single group as it varies in ability level, so as to suggest an unfair advantage for a higher-scoring group, while ignoring the proper counterpoise for that situation. That counterpoise is, of course, the improvement in hiring decisions that results from selecting higher scorers on valid tests. Comparisons of error rates between higher-scoring and lower-scoring groups, therefore, should be based on a method that reflects the trade-off between the errors within each group *before* comparing the groups themselves. This is what measures of validity, which weight the two kinds of error equally, are for (e.g., see the efficiencies in Fig. 1).

Figure 2A reveals details of the relations among single cells and marginals, and their comparability for both races, but a crucial and more direct demonstration of the artifactual nature of Seymour's more complicated error rates, based as they are on components from two cells, involves predicting those rates as he defined them via regression analysis from the variation of passing percentages within race for the 47 studies.

Regression analyses. Simple regression equations were derived for each race separately, with the percentage passing (Pass%) as independent variable, and with Seymour's acceptance and rejection errors as dependent variables ($N = 47$ for each race). Each mean passing percentage rate from Fig. 2B was then entered, as a value for the independent variable, in the two regression equations derived from data for the other race, to predict Seymour's two error rates had each race exhibited the Pass% of the other. If the predicted rates approximate the observed rates, so that the pattern emerging—as reflected in his racial ratios for each error rate—is the same as the one on which he bases his latest critique of tests, it will indicate that his racial contrasts imply nothing more than would contrasts between two samples from a single race that have passing rates equalling those of blacks and whites in Fig. 2B. In that event, his findings would have no bearing on race, but only on differences in passing rates.

Differences between the two regression coefficients for black and white equations predicting each of Seymour's two error rates are trivial and not statistically significant. For predicting his acceptance errors, the black and white regression coefficients were $b = .69$ and $b = .57$. (Deletion of one study's pair of outlying data, produced by a white sample with $n = 27$, would have raised the white coefficient to $b = .74$, and its accompanying r from .30 to .49.) For predicting Seymour's rejection errors, the black and white coefficients were $b = -.93$ and $b = -.91$. Correlation ratio and second- and third-degree polynomial tests for nonlinearity did not approach significance.

Since the regression lines can all be considered linear and also parallel within racial pairs for the same error rate, it is reasonable to extrapolate Seymour's error rates for each race using the regression equation of the other. Any shortfall in prediction could be due only to differences between the races in intercepts of the regression lines, and to random departures from perfectly parallel regression lines. Although using common regression coefficients would have been justified by the significance tests, separate coefficients were retained in order to examine the outcome more severely.

The unreliability of the tests used in the 47 studies and their sampling errors will be reflected to some degree in the unreliability of the test Pass% used as the independent variable in the regression analyses. Therefore, the planned regression analyses amount to

TABLE 1
Mean Acceptance and Rejection Error Rates Observed and Predicted Using Seymour's Definitions

	Acceptance errors[a]			Rejection errors[b]		
	Whites (%)	Blacks (%)	*t*	Whites (%)	Blacks (%)	*t*
Observed	50.5	32.6	8.2*	20.4	35.8	9.1*
Predicted	47.9[c]	37.9[d]	4.6*	15.3[e]	40.6[f]	14.9*

Note. For regression analysis, the observed and predicted rates must reflect the equal weighting of all 47 studies, regardless of sample size. Significance tests of differences between predicted means employ the same standard deviations as those of differences between corresponding observed means.

[a] Defined as FP/(FP + TN); the SD of this quantity and its correlation with Pass% were SD = 12.4 and r = .30 for whites, and SD = 8.4 and r = .78 for blacks.

[b] Defined as FN/(FN + TP); the SD of this quantity and its correlation with Pass% were SD = 6.3 and r = −.93 for whites, and SD = 9.8 and r = −.90 for blacks.

[c] Underprediction was expected from using black equation.

[d] Overprediction was expected from using white equation.

[e] Underprediction was expected from using black equation.

[f] Overprediction was expected from using white equation.

*p < .00001.

using regression equations derived for one group to predict for another group when the two groups differ greatly in predictor and criterion means and the predictor lacks perfect reliability. Even if regression lines are parallel, these conditions lead to underpredicting the criterion for a higher group (i.e., whites) when using the equation developed from a lower group (i.e., blacks) and to overpredicting the criterion for a lower group when using the equation developed from a higher group simply because of differences in intercept (Hunter & Schmidt, 1976; Jensen, 1980, pp. 512–514; Linn & Werts, 1971).

One can anticipate, therefore, that the pattern of predicted results will exhibit underprediction of Seymour's error rates when the black rates are determined from the equations for whites and overprediction when the white rates are determined from the equations for blacks. Moreover, the consequences of these anticipated effects will be exaggerated when comparing Seymour's observed racial ratios with the corresponding ratios based on the predictions, because the white/black ratio for predicted acceptance errors will have its numerator underpredicted and its denominator overpredicted and the black/white ratio for predicted rejection errors will have its numerator overpredicted and its denominator underpredicted.

Results in Tables 1 and 2 bear out these various expectations. Seymour's fundamental between-race effects are successfully reproduced from within-race data, and under- and overprediction appear where expected. In Table 1, the racial differences are significant at extremely small probabilities whether based on observed or predicted mean error rates, just as Seymour found for the former. In addition, the patterns of racial differences are identical for the observed and predicted error rates of both types, indicating that Seymour would have arrived at the same conclusion for both types of data, namely, that tests were unfair to blacks. However, each predicted set of data has been generated from only one race. Thus, while it may be true, as Seymour (1989) states, "that the probability that [these results] arose from chance is infinitesimally small," the probability that they arose simply

TABLE 2
Racial Ratios of Acceptance and Rejection Error Rates for Observed, Predicted, and Observed with Predicted Means

Error ratios	Obs./Obs.	Pre./Pre.	Pre./Obs.	Obs./Pre.
Acceptance, white/black	1.5	1.3[a]	1.5[b]	1.3[c]
Rejection, black/white	1.8	2.6[d]	2.0[e]	2.3[f]

Note. Based on the data in Table 1.
[a] Expected underprediction divided by expected overprediction.
[b] Expected underprediction divided by observed mean rate.
[c] Observed mean rate divided by expected overprediction.
[d] Expected overprediction divided by expected underprediction.
[e] Expected overprediction divided by observed mean rate.
[f] Observed mean rate divided by expected underprediction.

as an outcome of the differences between blacks and whites in right-hand marginal passing rates on the test seems certain. Chance is hardly the relevant hypothesis.

Table 2 compares the observed and predicted results from the standpoint of Seymour's other major analysis, of ratios. The racial ratios involving two predicted means are comparable to those that he relied upon. But basing ratios on a combination of predicted and observed means, as in the last two columns of the table, prevents the under- and overpredictions from the regression analyses from combining to exaggerate any discrepancies, and so those mixed results in Table 2 reveal an even greater degree of comparability with the entirely observed ratios than do the entirely predicted ratios. Once again, data derived from variation in ability within either race alone have successfully simulated Seymour's supposedly racial results.

Analysis based on interchanged right-hand marginals. Lest it be thought that the minor deviations from perfect fit in Tables 1 and 2 contain any residuum of support for Seymour's (1988, 1989) conclusions, perhaps stemming from minor differences between blacks and whites in the conventional True Positive and True Negative rates, one final demonstration of the effect of differences in passing percentages is offered, based entirely on the variation of those percentages between races, as distinct from within races. For this analysis, the right-hand marginal distributions of blacks and whites in Fig. 1 have been *interchanged* with each other, leaving in place the True Positive and True Negative rates (and, of course, the False Positive and False Negative rates), which define and reflect the predictive values of the test (Galen & Gambino, 1975). Seymour's error rates were then calculated from cell frequencies generated by the resulting imposition of the test passing rates of each race on the predictive values of the other. Those error rates are then compared, in Table 3, with the observed error rates from the aggregate data for each race in Fig. 1 (the appropriate reference, since both sides of the comparison involve the weighting of the 47 studies by sample size).

To avoid confusion in presenting results of this analysis, and for convenience in arranging Table 3, the interchanged fourfold tables are referred to by the racial label originally associated with the right-hand marginals. Thus, for the results of the interchanged analyses, "whites" refers to the combination of white Pass% marginals with black True cell rates and their False cell complements, and, similarly, "blacks" refers to the combination of black right-hand marginals with white True cell interiors. Readers who, understandably, may prefer to regard the True cell interiors as defining the essential identity of the interchanged tables need only think "blacks" where Table 3 lists "whites" and vice versa.

TABLE 3
Acceptance and Rejection Errors and Their Racial Ratios for Observed and Interchanged Right-Hand Marginals (Percentages Passing)

	Acceptance errors[a]			Rejection errors[a]		
	Whites (%)	Blacks (%)	W/B	Whites (%)	Blacks (%)	B/W
Observed	54.8	33.4	1.6	20.4	37.1	1.8
Interchanged	57.4	31.2	1.8	18.0	40.7	2.3
Obs./Inter.	54.8	31.2	1.8	18.0	37.1	2.1
Inter./Obs.	57.4	33.4	1.7	20.4	40.7	2.0

Note. The observed and interchanged rates are based on the aggregate data of Fig. 1, to which the 47 studies contribute according to their sample sizes. In the interchanged analyses, the racial identity of samples is defined in the table by the race to which the marginals originally belonged. Thus, "whites" refers in those analyses to the combination of the white Pass% marginals with the black True Positive and True Negative rates, and "blacks" refers to the combination of black Pass% marginals with the white True Positive and True Negative rates.

[a] Seymour's definitions.

Table 3 shows that a simple interchange of white and black right-hand marginals, without altering the intrinsic True cell rates of either race, produces error rates as defined by Seymour that are consistently even more favorable to the group with the higher passing rate than those that he observed. This conclusion is most accessible simply from comparing the white/black and black/white ratios in Table 3. Of greater interest, perhaps, are the racial ratios produced by mixing interchanged with observed error rates, because these ratios combine the interiors of just one race with the right-hand marginals of both; these mixed comparisons, too, consistently produce racial ratios (and racial differences) that are more favorable to the higher-scoring group than those that Seymour observed, *even though, in every such comparison, both fourfold tables have exactly the same True Positive and True Negative rates and differ only in marginal passing rates.* The interchanged analyses produce even more extreme outcomes than Seymour observed because, by holding True cell rates constant, they fail to reflect the adjustments that those rates typically undergo in real data in response to different right-hand marginals that offset to some extent the effects of those marginals on sensitivity and specificity (and hence on Seymour's error rates).

Thus, separate empirical analyses based on (a) within-group variation and (b) between-group variation demonstrate that Seymour's effects are simply reflections of differences in levels of tested ability and have nothing, fundamentally, to do with race.

REFERENCES

Alderfer, C. P. (1982). Reflections on polarities and bias. In A. S. Glickman (Ed.), *The changing composition of the workforce: Implications for future research and its application* (pp. 145–151). New York: Plenum.

Allen, W. B. (1988). Rhodes handicapping, or slowing the pace of integration. *Journal of Vocational Behavior,* **33,** 365–378.

Beeghley, L., & Butler, E. W. (1974). The consequences of intelligence testing in the public schools before and after desegregation. *Social Problems,* **21,** 740–754.

Beer, W. R. (1987). Resolute ignorance: Social science and affirmative action. *Society,* **24,** 63–69.

Bolick, C. (1988). Legal and policy aspects of testing. *Journal of Vocational Behavior,* **33,** 320–330.

Cleary, T. A. (1968). Test bias: Predicting grades of Negro and white students in integrated colleges. *Journal of Educational Measurement,* **5,** 115–124.

Churchman, C. W. (1961). *Prediction and optimal decision: Philosophical issues of a science of values.* Englewood Cliffs, NJ: Prentice–Hall.

Davis, B. D. (1986). *Storm over biology: Essays on science, sentiment, and public policy.* Buffalo: Prometheus Books.

Davis, J. A. (1971). *Elementary survey analysis.* Englewood Cliffs, NJ: Prentice–Hall.

DuBois, W. E. B. (1903). The talented tenth. In B. T. Washington, W. E. B. DuBois, P. L. Dunbar, & C. W. Chesnutt, et al., *The Negro problem.* New York: James Pott.

Dunnette, M. D. (1974). Personnel selection and job placement of disadvantaged and minority persons: Problems, issues, and suggestions. In H. L. Fromkin & J. J. Sherwood (Eds.), *Integrating the organization: A social psychological analysis* (pp. 55–74). New York: Free Press.

Galen, R. S., & Gambino, S. R. (1975). *Beyond normality: The predictive value and efficiency of medical diagnoses.* New York: Wiley.

Gordon, R. A. (1977). A critique of the evaluation of Patuxent Institution, with particular attention to the issues of dangerousness and recidivism. *Bulletin of the American Academy of Psychiatry and the Law,* **5,** 210–255.

Gordon, R. A. (1980a). Examining labelling theory: The case of mental retardation. In W. R. Gove (Ed.), *The labelling of deviance* (2nd ed., pp. 111–174). Beverly Hills, CA: Sage Publications. (Original work published 1975)

Gordon, R. A. (1980b). Labelling theory, mental retardation, and public policy: *Larry P.* and other developments since 1974. In W. R. Gove (Ed.), *The labelling of deviance* (2nd ed.) (pp. 175–225). Beverly Hills, CA: Sage Publications.

Gordon, R. A. (1982). Preventive sentencing and the dangerous offender. *British Journal of Criminology,* **22,** 285–314.

Gottfredson, L. S. (Ed.). (1986a). The *g* factor in employment [Special issue]. *Journal of Vocational Behavior,* **29,** 293–296.

Gottfredson, L. S. (1986b). Societal consequences of the *g* factor in employment. *Journal of Vocational Behavior,* **29,** 379–410.

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

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

Hawk, J. A. (1970). Linearity of criterion–GATB aptitude relationships. *Measurement and Evaluation in Guidance,* **2,** 249–251.

Homans, G. C. (1984). *Coming to my senses: The autobiography of a sociologist.* New Brunswick, NJ: Transaction Books.

Horowitz, I. L. (1987). Disenthralling sociology. *Society,* **24,** 48–55.

Hunter, J. E. (1983). *Fairness of the General Aptitude Test Battery: Ability differences and their impact on minority hiring rates* (United States Employment Service Test Research Report No. 46). Washington, DC: Division of Counseling and Test Development, Employment and Training Administration, U.S. Department of Labor.

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

Hunter, J. E., & Schmidt, F. L. (1982). Fitting people to jobs: The impact of personnel selection on national productivity. In M. D. Dunnette & E. A. Fleishman (Eds.), *Human performance and productivity: Human capability assessment* (Vol. 1, pp. 233–284). Hillsdale, NJ: Erlbaum.

Jensen, A. R. (1974). How biased are culture-loaded tests? *Genetic Psychology Monographs,* **90,** 185–244.

Jensen, A. R. (1977). An examination of culture bias in the Wonderlic Personnel Test. *Intelligence*, **1**, 51–64.

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

Larry P. v. Riles, 495 F. Supp. 926 (N.D. Cal. 1979).

Lewis, M. A. (1988). *Detailed critique of the Seymour article* (*Tests which work differently*). Unpublished manuscript.

Linn, R. L., & Werts, C. E. (1971). Considerations for studies of test bias. *Journal of Educational Measurement*, **8**, 1–4.

Lipset, S. M. (1981). The limits of social science. *Public Opinion*, **4**, 2–15.

Lipset, S. M., & Ladd, E. C., Jr. (1972). The politics of American sociologists. *American Journal of Sociology*, **78**, 67–104.

Mercer, J. R. (1973). *Labeling the retarded*. Berkeley: University of California Press.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist*, **35**, 1012–1027.

Mitroff, I. I., & Sagasti, F. (1973). Epistemology as general systems theory: An approach to the design of complex decision-making experiments. *Philosophy of Social Science*, **3**, 117–134.

Padover, S. K. (Ed.) (1953). *The complete Madison: His basic writings*. New York: Harper.

Peckham, R. F. (1979). [Opinion, *Larry P. v. Riles*, 495 F. Supp. 926 (N.D. Cal. 1979)], United States District Court, Northern District of California.

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

Quigley, A. M. (1988). *Ex nihilo nihilo: A reply to Seymour*. Unpublished manuscript.

Reynolds, H. T. (1977). *The analysis of cross-classifications*. New York: Free Press.

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., & Hunter, J. E. (1980). The future of criterion-related validity. *Personnel Psychology*, **33**, 41–60.

Schmidt, F. L., Hunter, J. E., McKenzie, R. C., & Muldrow, T. W. (1979). Impact of valid selection procedures on workforce productivity. *Journal of Applied Psychology*, **64**, 609–626.

Schreier, F. L. (1957). *Human motivation: Probability and meaning*. Glencoe, IL: Free Press.

Seymour, R. T. (1988). Why plaintiffs' counsel challenge tests, and how they can successfully challenge the theory of "validity generalization." *Journal of Vocational Behavior*, **33**, 331–364.

Seymour, R. T. (1989). Tests which work differently between blacks and whites: The Achilles' heel of "validity generalization." In B. Gifford (Ed.), *Testing and public policy*. Norwell, MA: Cluwer Academic Publishers. (Available from the Lawyers' Committee for Civil Rights under Law, Suite 400, 1400 Eye Street, Northwest, Washington, DC 20005.)

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

Simpson, M. (1980). The sociology of cognitive development. *Annual Review of Sociology*, **6**, 287–313.

Society for Industrial and Organizational Psychology, Inc. (1987). *Principles for the validation and use of personnel selection procedures* (3rd ed.). College Park, MD: Author.

Spitz, H. H. (1986). *The raising of intelligence: A selected history of attempts to raise retarded intelligence*. Hillsdale, NJ: Erlbaum.

Stanley, J. C. (1957). Index of means vs. mean of indices. *American Journal of Psychology*, **70**, 467–468.

United States Employment Service (1970). *USES (1970) manual for the USTES General*

Aptitude Test Battery: Section III. Development. Washington, DC: U.S. Government Printing Office.

Wainer, H. (1987). *Can we accurately assess changes in minority performance on the SAT?* (Technical Report No. 87-75). Princeton: Educational Testing Service.

Watkins, B. T. (1982, September 22). Sociology 101: "Unduly eclectic, unscientific, thin in substance." *The Chronicle of Higher Education,* p. 6.

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.

Zelditch, M., Jr. (1959). *A basic course in sociological statistics.* New York: Holt, Rinehart & Winston.

Zeisel, H. (1957). *Say it with figures* (rev., 4th ed.). New York: Harper & Row.

Received: October 19, 1988