BAD NEWS CONCERNING IQ TESTS

ROBERT A. GORDON AND EILEEN E. RUDERT
Johns Hopkins University

Sociology of Education 1979, Vol. 52 (July): 174–190

This issue's paper by Guterman shows that IQ predicts important criterion variables equally well across social class in a predominantly white sample, leading to the conclusion that IQ tests are not socioeconomically biased. The present paper places Guterman's finding in the context of more general studies concerned with possible racial bias in IQ tests. These studies approach the problem of bias in two major ways, through external validity, like Guterman's, and through internal validity, which deals with the detailed behavior of test items themselves. Both kinds of study agree in showing that IQ tests are not biased against blacks. In addition, studies of internal validity show that the model of test bias embraced by test critics implies a process of cultural diffusion that is improbable. Questions raised by Guterman concerning the legitimacy of using IQ in stratification research are discussed. Finally, IQ is compared with SES in status attainment models for blacks and whites, where it is shown to play roughly the same role for both races, and to have generally stronger direct effects than SES throughout the models.

In this journal issue, Guterman (1979) has examined the predictive or construct validity of an IQ test across social class and found no evidence that the test artificially understates the pragmatic intelligence of members of lower strata as that intelligence is normally reflected in socially significant criterion performances, some of which were measured at different points in time. The IQ test employed was the Quick Test, which is based on recognition vocabulary (Ammons and Ammons, 1962). Vocabulary tests have consistently proven to load strongly on the general factor, *g*, which runs through all cognitive tests, and which is commonly referred to as "intelligence." Although early in its history the word "intelligence" was understood to refer more or less directly to innate ability (Burt, 1970), as sophistication has increased concerning the crucial distinction between genotype and phenotype, users now often restrict its meaning to phenotypic intelligence (e.g., Jensen, 1969:19–20) unless they explicitly state otherwise. Since phenotypic values can be measured directly, within the limits of measurement error, whereas genotypic ones must be inferred indirectly through heritability analyses, the more contemporary usage renders the often repeated truism that "intelligence cannot be measured directly" somewhat ambiguous. Obviously, in the phenotypic sense the truism is not true.

All of Guterman's criteria but one represent variables that are understood to depend on IQ without simply being alternative versions of IQ tests themselves. The lone exception is the GATB–J, also a vocabulary test of ability, but with a somewhat stronger reasoning component than the Quick Test since the former requires the pairing of synonyms and antonyms. The GATB–J is one of the three subtests in its battery, and the only verbal one, used to measure general intelligence (Blum, 1953:687; Dvorak, 1956). As used in its present role it adds little to Guterman's argument, consequently, and blurs the distinction between aptitude or IQ measures on the one hand and achievement measures on the other hand that is embodied throughout the rest of his paper. More mileage might have been gotten out of this instrument by employing it as a replication on the independent variable side of the analysis.

Guterman's other dependent variables fall nicely into three main groups, however, that in combination lend a special strength to the design of his study. One group consists of more or less standard achievement tests, the second of grade point average at three different points in time, and the third of various kinds of knowledge and information tests whose content is much less closely tied to usual school curricula than that of the achievement tests in the first group. These groups

of variables sample the full range of cognitive performance in respect to IQ with which society is usually explicitly concerned, since the achievement tests and grade point averages can be viewed as being somewhat analogous to measures of performance in job training programs, the only major domain omitted in Guterman's study.

Guterman finds that a composite index of socioeconomic status (SES) based on father's occupation and each parent's education adds less than one percent of variance, net of IQ, to reading and arithmetic achievement, to self–reported grade point average (GPA) in 9th, 10th, and 11th grades, and to two measures closely related to occupational knowledge ("Job Information" and "Military Knowledge"). "Political Knowledge" shows an increment of about one percent, and "Sexual Knowledge" an increment of four percent. Since the correlation between SES and the somewhat bookish Sexual Knowledge test is a robustly positive .39, which was higher than that for any other knowledge test (Guterman, 1979:Table 3), we can probably surmise that in this one area at least practical experience is not a total substitute for "book–learning."

None of this is particularly surprising in view of overwhelming published evidence that one of us has reviewed elsewhere (R. Gordon, 1975:91–102) showing that IQ tests are not unfairly biased against the even more severely disadvantaged black population in the United States. Certainly, there existed a much stronger prima facie case for test bias across race than for test bias across class within the white population. As one of us (R. Gordon, 1976:260) noted elsewhere after reviewing the evidence concerning race and citing other concurring reviews, "Given these facts, it follows a fortiori that socioeconomic bias in the meaning of IQ scores within the white population itself poses no problem of any consequence." Guterman has now shown this conclusion to have been correct.

It is surprising that Guterman has not drawn more extensively on the studies of test fairness with respect to other groups called "culturally disadvantaged," in view of his reference to this literature at several points, especially since that literature strongly bolsters his own findings. Although he expressed concern over the "legitimacy" of using IQ data in status attainment research, it is also the case that many such studies not only include blacks, but also compare the status attainment process for blacks and whites. Guterman does not even inform readers that 256 blacks were present in the national sample of 2,213 tenth–grade boys that he used (Bachman, 1970:5), although this would seem to make his results more general. Let us compensate for these omissions, therefore, by briefly recapitulating some of the major findings concerning test bias across race from an earlier review (R. Gordon, 1975) and, where necessary, noting developments that have occurred since the time at which it was written. For simplicity, and because of its special importance to many recent sociological studies, we now confine ourselves largely to the evidence concerning blacks. Readers interested in findings concerning Mexican–Americans, whose situation incorporates the additional complexity of bilingualism, may wish to refer to the earlier review, where they will find that the evidence supports a conclusion essentially similar to that reached for blacks.

TEST FAIRNESS AND RACE

Studies of External Validity

Studies of test fairness fall under two major methodological headings, one of which is "external validity." Guterman's (1979) analysis is an example of this method. It differs superficially in appearance from other examples concerning race because Guterman's background variable, SES, can be treated as a continuous variable, unlike race. The logic, however, is the same—that of moderated prediction, where group membership serves as the moderator variable.

Studies of external validity are usually prompted by issues of fairness that arise in connection with selection for admission to college, graduate school, or professional school. Selection is accomplished by regressing GPA on aptitude test scores and other variables such as high school class

rank, and using the resulting equation for ranking new applicants. Typically, there are too few blacks in selective institutions to warrant applying a separate equation to them even if their small numbers permitted finding a stable one. Consequently, the separate question of whether or not the equation for white or mostly white samples is fair to blacks also emerges. A favorable answer concerning this last question is regarded as a favorable answer to the issue of overall fairness, whereas a negative answer here would still leave open the possibility that a separate equation for blacks would prove satisfactory.

Investigators are in good agreement that tests predict grades for blacks that are as high as or higher than the grades actually obtained. That is, when tests are used to select blacks on the basis of predicted grades by employing regression equations developed from either entirely white or mixed samples, they tend to exhibit bias in favor of blacks. This phenomenon is known as "overprediction," in contradistinction to the "underprediction" that was expected by those alleging bias in the tests. In statistical terms, the regression slopes are usually similar for both races, but blacks often have a lower intercept. For other reviews and specific empirical studies in academic contexts, the reader is referred to Boney (1966), Stanley and Porter (1967), Hills and Stanley (1968, 1970), APA Task Force on Employment Testing of Minority Groups (1969), Thomas (1971), Stanley, (1971a, 1971b), Cleary, Humphreys, Kendrick, and Wesman (1975), Lerner (1976), and Goldman and Hewitt (1976). Additional sources that specifically display the overprediction phenomenon for minority groups are M. Gordon (1953), Cleary (1968), Kallingal (1971), Pfeifer and Sedlacek (1971), Temp (1971), Goldman and Richards (1974), R. Gordon (1975:Table 4.3) using data from the Coleman Report (Coleman, Campbell, Hobson, McPartland, Mood, Weinfeld, and York, 1966), and Linn (1975).

The demonstrated validity of intelligence tests for blacks in the area of academic performance would naturally be expected to generalize most strongly to those occupational performances in which abstract reasoning plays a conspicuous role, such as the professions. However, points in the occupational range as far removed from the professions as clerk–typist and teletype operator (M. Gordon, 1953), machine–shop employee in the aircraft industry (Tenopyr, 1967, after Thomas, 1971:73), medical technician in Veterans Administration hospitals (Campbell, Flaugher, Pike, and Rock, 1969), telephone company installation and repairman (Grant and Bray, 1970), and telephone company service representative (Gael and Grant, 1972) have now been studied. Again, tests prove equally valid for blacks and whites, with any bias due to overprediction favoring blacks (M. Gordon, 1953; Campbell et al., 1969; Grant and Bray, 1970). These general findings extend to the broad subsociety of military occupations (M. Gordon, 1953; Maier and Fuchs, 1975).

External validity can also be examined in purely correlational studies employing the usual validity coefficient. Occasional differences between blacks and whites in such studies (e.g., Dalton, 1974) probably reflect artifacts such as restriction of range due to the fact a test was too difficult—that is, lacking sufficient "floor." This possibility was demonstrated to be real by Hills and Stanley (1968, 1970), who showed that the level of the School and College Abilities Test (SCAT) appropriate for grades 6 to 8 predicted freshman grades better than did the Scholastic Aptitude Test (SAT) at three predominantly black southern colleges.

Practically all other correlational studies involve employment tests, and fall into one of two categories, depending on whether the validity correlations for each race are independently compared with zero or are compared directly with each other (Humphreys, 1973). Those in the first category are sometimes referred to as "single–group validity" studies, those in the second as "differential validity" studies, although differential validity is, of course, the question at issue in both categories. Schmidt, Berner, and Hunter (1973) examined nineteen studies reporting correlational validities separately by race and found that a statistical model that assumes no true differences between races, but which takes account of dif-

ferences in black and white sample sizes and general level of test validity, could adequately account for the observed data. They expressed doubt, in view of their results, that the phenomenon of single–group validity existed, a view shared by Humphreys (1973). Essentially identical results concerning single–group validity have been obtained from studies of the cumulative research by O'Connor, Wexley, and Alexander (1975), Katzell and Dyer (1977), and Boehm (1977). In a later review of 39 studies, Hunter, Schmidt, and Hunter (forthcoming) have demonstrated that findings of apparent differential validity can also be accounted for by chance and certain statistical artifacts stemming from the requirement by previous reviewers that at least one of the validities in a black–white validity pair be statistically significant in order for the pair to be included in their analyses. Differential validity, Hunter et al. conclude, is probably a nonexistent phenomenon, too. In short, there is no evidence that tests are valid for one race but not the other, or that tests are more valid for one race than the other. It might also be added that these conclusions applied equally well regardless of whether the criterion measures of job performance were subjective or objective.

Studies of Internal Validity

Studies of "internal validity," the second major method of investigating test fairness, seek evidence concerning cultural bias by examining the internal structure of the tests themselves. If bias is present, it should show up in group–by–item interactions when tests are administered to different populations. Such interactions can be studied as a single component in appropriate analysis of variance models, and they can also be decomposed for finer analysis into two more basic kinds of evidence. One kind consists of significant differences between groups in the rank order of difficulty of items (as indicated by the percentage, p, passing each item), and the other "is seen even when the rank order of p values is the same in both groups but the *differences* between the p values of adjacent items are significantly different in the two populations" (Jensen, 1974:189). Jensen has been the most active person in this line of research, comparing tests that vary considerably in their prima facie "cultural loading" such as the nonverbal Raven's with the highly verbal Peabody Picture Vocabulary Test, across white, black, and Mexican–American populations.

Even today, it is often the case that the face validity of particular items ("How many blacks will have seen an artichoke?") and the manifest cultural loading of some tests are adduced, in conjunction with mean differences between groups, as conclusive evidence of test bias. Mercer (1978:1664), for example, explicitly cited what she termed "Anglo" cultural content of items, mean differences, and the ability to reduce a difference by controlling statistically for background as the test–related evidence she relies on to determine bias, in her testimony concerning tests before a U.S. District Court. Indeed, if Flaubert were alive to revise his mordant *Dictionary of Accepted Ideas*, one would surely be able to look up "Intelligence tests" and find "Culturally biased." More effectively than the practical operationalism of the regression approach, the group–by–items interaction method reveals why this form of argument—which clearly begs the question as to the meaning of the differences—is psychometrically naive.

If group differences are due entirely to differential exposure to information required by particular items, those particular items should become differentially difficult for the disadvantaged group, and this special difficulty should be reflected by detectable changes in the rank order of difficulty of the affected items. By excluding such items from an instrument, it should be possible to construct a test that preserves the construct validity of IQ within race while eliminating the mean difference between races. Conceivably, judicious choice of items should even enable one to construct a test that reverses the direction of the usual mean difference.

The much publicized BITCH test, based on recognition of black slang, appears to achieve this last objective (Williams, 1972), but there is no evidence

that it functions as an intelligence test even among blacks (Humphreys, 1975). Those of its items that we have seen do not employ the analogical reasoning format frequently found in higher level verbal tests such as the SAT Verbal Aptitude (SAT–V), part of which asks, "A is to B as C is to . . . D, E, F, G?" thus requiring the understanding of at least four words in order to get credit for a correct answer. The difficulty one would encounter in attempting to frame items in this format with slang suggests just how restricted the domain is that the BITCH test draws upon.

In his studies of the Raven's Colored Progressive Matrices and the Peabody, Jensen found that the total interaction of ethnic–group–by–items accounted for "an exceedingly small proportion of the total variance" in either test (1974:241). One of the more elegant and theoretically meaningful analyses that Jensen performed involved the demonstration that even the modest group–by–items interaction that had been observed could be eliminated almost entirely when the analysis of variance was based on children from a minority group and a white group about one or two years younger than the minority group. This shows that the group–by–items interaction can be interpreted as a mental–age–by–items interaction rather than as a cultural–differences–by–items interaction. Pairing a younger white group with a lower IQ ethnic group equates the two for mental age. Jensen then bolstered this argument concerning mental age as the main source of the slight interaction by repeating the analyses with two groups of whites, one group selected so as to average two years older than the other. This enabled him to show that with the younger whites thus simulating a "pseudo–ethnic" group, an interaction of similar magnitude could be elicited, but this time using groups from the same culture, differing only in mental age.

Identical results were obtained by Jensen (1977) for the Wonderlic Personnel Test, which, compared with the Peabody and Raven's, is of intermediate cultural loading. Since this study involved adults, Jensen accomplished his "pseudo–ethnic" pairing by working inward from the two ends of the white score distribution so as to create two groups of whites one standard deviation apart in mean score. The "race"–by–items interaction here was comparable to that between unselected blacks and whites, and much larger than that observed between blacks and whites when they had been matched for total scores, i.e., level of ability. As a final test of the method of determining cultural bias by inspection of face validity, Jensen submitted the eight most and eight least racially discriminating Wonderlic items to two groups of psychologically sophisticated judges, five blacks and five whites. They were informed of the nature of the sixteen items, and requested to sort them into their original categories solely by inspection. Neither group attained even the chance level of accuracy, and only one individual exceeded it at all.

The Wonderlic has been used by more than 6,500 business organizations as part of their personnel selection and placement procedures. It should be of special interest to students of status attainment that the median and mean test score correlations across 80 occupational categories between blacks and whites are .84 and .87, respectively (Jensen, 1977:53), thus indicating a high degree of similarity in the way blacks and whites apply for jobs in relation to their ability.

Jensen (1976) has also described studies of the rank order of difficulty of items from the Stanford–Binet and Wechsler Intelligence Scale for Children (WISC). Nichols (1970:Tables 13–15) analyzed the sixteen most heterogeneous items of the Stanford–Binet, i.e., those most likely to display interactions, and it was found that the rank order of difficulty correlation between black and white preschoolers was .99. Jensen noted that this occurs even before the children have been exposed to the common culture of the school. Over all 161 WISC items, the cross–racial order–of–difficulty correlation was .95; when white children were compared with blacks two years older, slight disparities in rank were reduced even further, rather than enlarged, as revealed by a correlation of .96. (See also Miele, 1979.)

One WISC item, above all, deserves special comment, since it has so often been singled out by test critics as a self–

evident example of glaring bias: "What is the thing to do if a fellow (girl) much smaller than yourself starts to fight with you?" According to black psychologist Williams, on the "CBS Reports" program, "The IQ Myth," "in our society—black communities—a child is told that, if another child hits him, hit him back" (CBS, 1975:4). WISC author Wechsler himself, on the same program, conceded the item might require different scoring for blacks, saying, "we had protests, especially from people in underdeveloped areas . . . they said the thing to do is to bust him in the jaw" (CBS, 1975:4). These remarks prompted a student to examine the item statistics (Jensen, 1976:343–344), and it was found that this item ranked 42 out of 161 for black children as compared to 47 for whites, when the easiest item was ranked first. That is, the item was not particularly difficult for either race and, relative to other items, easier for blacks. Dr. Jerome Doppelt, who actually directed the projects for standardizing both the original WISC and the recently revised WISC–R (for which the standardization sample included a representative proportion of blacks), and who is now Director of the Psychological Measurement Division of the Psychological Corporation, was questioned closely about this same item in a deposition taken in connection with *Larry P.*, a civil complaint brought against the use of IQ tests in California public schools. Doppelt (1977:45–48) stated that he had looked into the behavior of this item in the WISC–R standardization sample, and found that about 81 percent of the nonwhite children answered correctly compared to about 86 percent of the whites. In the age-range from 8.5 to 10.5, the percentage for nonwhites answering correctly was actually "a little bit higher than whites." Doppelt stated that on the basis of these statistics—which, like Jensen's, show the item to be an easy one for both races with no unusual difference between them—he would score the item according to the manual, Wechsler's televised remarks notwithstanding.

The statistical facts with respect to this favorite example of test critics epitomize the absence of (specifically) race–by–item interaction in studies of internal validity. It should be emphasized that the method is a powerful one, and that when questionable items are present it is quite capable of detecting them. Medley and Quirk (1974) found such an interaction between race and the content of items in the general culture section of the National Teacher Examination, which at one time had overlooked black artists and social problems related to race, and Bhushan (1974) has presented illuminating specimens of faulty items that were uncovered in the course of translating an IQ test from English into French. Cleary and Hilton (1968) detected a trivial amount of interaction in a few Preliminary Scholastic Aptitude Test (PSAT) items, but Stanley (1969) demonstrated that it was an artifact due to the extreme difficulty of a few items for members of both races. Angoff and Ford (1973) continued analysis of the PSAT for blacks and whites and found little evidence of bias. By means of using controls similar to those employed by Jensen, they concluded that the small amount of interaction observed was better attributed to population differences in level of ability than to cultural differences between races. Chase and Pugh (1971) employed the same general technique to search for bias in an intelligence test given to middle class and lower class white children, thus making their study directly relevant to Guterman's one of external validity. Again, no evidence of bias was found.

The fact that there is a slight level–of–ability–by–items or mental–age–by–items interaction in many of these studies can be understood if one considers that items are not all equally saturated with g. Consequently, the order of difficulty of items is shifted slightly when comparisons are made between groups that differ appreciably in mental age or IQ. When the groups are different races, the slight interaction from this source becomes confounded with one potentially due to cultural differences. However, simulation of the interaction using groups from the same race but differing in ability in studies by Jensen and Angoff and Ford reveals its true nature.

The absence of race–by–item interaction in all of these studies places severe

constraints on models of the test score difference between races that rely on differential access to information. In order to account for the mean difference, such models must posit that information of a given difficulty among whites diffuses across the racial boundary to blacks in a solid front at all times and places, with no items leading or lagging behind the rest. Surely, this requirement ought to strike members of a discipline that entertains hypotheses of idiosyncratic cultural lag and complex models of cultural diffusion (e.g., "two-step flow of communication") as unlikely. But this is not the only constraint. Items of information must also pass over the racial boundary at all times and places in order of their level of difficulty among whites, which means that they must diffuse across race in exactly the same order in which they diffuse across age boundaries, from older to younger, among both whites and blacks. These requirements imply that diffusion across race also mimics exactly the diffusion of information from brighter to slower youngsters of the same age within each race. Even if one postulates a vague but broad kind of "experience" that behaves in exactly this manner, it should be evident that it would represent but a thinly disguised tautology for the mental functions that IQ tests are designed to measure. As Jensen (1977:63) notes, "The only way one could view these findings as being consistent with the hypothesis that the Wonderlic is a culturally biased test would be to claim that culture bias depresses blacks' performance on all the test items to much the same degree." Thus, the critic who retreats into this position would find himself essentially in agreement with the conclusion that for most purposes for which IQ tests are used, a black with a given IQ is statistically indistinguishable from a white with the same IQ, not only in external performance, but also in the minute details of test performance itself.

Some readers may suppose that the model of cultural bias in the minds of the more formidable critics of IQ tests is actually more sophisticated than the one placed in question by studies of internal validity, and hence that it is capable of withstanding the challenge that such studies pose. Let us therefore examine the model advanced by one of the more prominent figures in the debate over the meaning of differences in IQ, Leon Kamin. During his testimony as an expert witness for the plaintiffs in *Larry P.*, Kamin was asked by the judge what he meant by "cultural bias." In what follows, readers should make allowances for rough spots in the court reporters' daily transcript. Kamin (1977:930) responded to the judge (emphasis added):

> The very fact that the tests must depend upon the particular information that a child has acquired in his past means that they are bound to be culturally biased. In different social classes in our society, in different ethnic groups in our society, in different racial groups in our society, the experiences which a child has vary. Now, the tests, for the most part, have been designed by white middle class psychologists who are familiar with white middle class environment, white middle class culture and understand what it is that one learns and acquired in that environment. And quite naturally, I believe, they have drifted into taking the source [sorts?] of *bits of information* and knowledge that their own children acquire, often from the parents, as an indication of "Ah ha, this is what an intelligent child ought to know." And when that is applied mechanically to children from very different backgrounds, either ethnically or racially or in terms of social class, it seems to me that a great bias is involved. Obviously, the children from other backgrounds will not have had the same access to, and the same experience with, *the bits of knowledge* tested on these tests as the modal white middle class child.

Note that it is precisely Kamin's "bits of knowledge" model that is rendered most implausible by the constraints imposed on any cultural diffusion process by the studies of internal validity. In order to accept Kamin's model, we must believe that "bits of knowledge" as divergent from each other as items on the nonverbal Raven's are from the vocabulary items on the Peabody, and as Performance items are from Verbal items on the WISC, diffuse across group boundaries in solid waves of equal difficulty, such that items of similar level of difficulty from tests of highly dissimilar content remain more closely linked with each other than with items of different difficulty but similar content from the same test. In short, we

must be willing to believe that information and content are simply one–dimensional for purposes of cultural diffusion, and that that single dimension just happens to coincide with age–graded difficulty. Apparently, the critics themselves must find this a highly implausible state of affairs, for Kamin made no effort in his testimony to confront the evidence from studies of internal validity, although several were in print by 1977 (Jensen, 1974, 1976; Angoff and Ford, 1973), and although Kamin has shown elsewhere (1974) that he follows Jensen's work closely.

One of the earliest explorations of group–by–items interaction in the literature concerns sex differences in response to certain items on the Stanford–Binet (Terman and Merrill, 1937:34; McNemar, 1942:45–54). The fact that items were eliminated or balanced so as to reduce sex differences in total IQ was a recurrent exhibit in evidence of cultural bias in the course of *Larry P.* Kamin (1977:875–876), for example, has stated:

> I am struck at the discrepancy in the treatment of the sex difference versus the treatment of the race difference and the treatment of the social class difference for that matter. I can see no scientific ground why one should eliminate questions which appear to show that one sex is doing better than the other and not eliminate questions which appear to show that one social class is doing better than another or that one race is doing better than the other.
>
> It seems to me this has to reflect the preconception of the people who are making up these tests. It seems very obvious that Terman and Merrill had a preconception that girls and boys were of equal intelligence. Therefore, when their items which they thought measured intelligence showed a large difference between males and females, they concluded these items just don't seem to be doing the job, and they removed them.
>
> Evidently, when they found the items discriminated between blacks or whites or between upper and lower-class people . . . I would imagine that the testmakers felt that these differences simply validated the test as a test of intelligence.

In this passage, Kamin is actually citing an exception that proves the rule and then reveals it to be a sound rule. The differences between the sexes were peculiar to just a small number of items, about 30 altogether (McNemar, 1942:50). About half the differences went in one direction and about half in the other. The great bulk of items *showed no differences*. It is precisely this varying nature of the group difference that constitutes interaction, and it is precisely this interaction that is lacking, according to the studies of internal validity, in race and social class comparisons. In the case of the sexes, there were many items showing no difference that could be retained, so that one would still have an intelligence test. The search for a set of items that would accomplish this for race and social class has been a diligent one, and it has failed to produce an adequate test. If one were to eliminate items showing a difference between the races or classes, as Kamin disingenuously recommends, all of the items would have to go.

Like Kamin, IQ test critic Mercer also "treats all psychological tests as measures of learned behavior" reflecting an "Anglo-Saxon cultural tradition" (Mercer and Lewis, forthcoming: 30, 12; see also Mercer, 1973; and R. Gordon, 1975). During her testimony for plaintiffs in *Larry P.*, Mercer was asked by plaintiffs' counsel, "Does Dr. Jensen have his own definition of bias?" to which she responded, "Yes. Recently Dr. Jensen has proposed a definition of test bias which is different from either customary usage or the statistical usage" (Mercer, 1977:1704). She went on to acknowledge that Jensen's term "culture loaded" corresponded to what she meant by "culture biased," and she gave an accurate description of the operationalization of the concept of bias in terms of group-by-items interaction. However, when asked, "Do you know of any evidence of studies that have been done of tests that show that in fact this definition tells us anything about cultural bias?" Mercer (1977:1708) responded:

> Well, I am not aware of any studies that show that this theoretical definition exists any place in the real world—it may—but that you would get this situation where the items would jump around for one group—one that was easy for one group would be hard for another group—evidence that this, in fact, ever occurs, I mean, I'm not aware of.
>
> My own hypothesis would be it would be very unlikely that this situation would ever occur in the real world, because, you see, the difficulty level of an item is determined

> by the cultural base from which the items are drawn. . . .

These extracts demonstrate clearly that major test critics not only adhere to, but wholeheartedly embrace, the model of diffusion of test information that looks so implausible in the light of studies of internal validity. In this testimony, Mercer could have mentioned the classic example of sex-by-items interaction in the Stanford-Binet presented by fellow-witness Kamin, the study by Medley and Quirk (1974) that found content bias in a section of the National Teacher Examination, and the account by Bhushan (1974) of bias in particular items created by translating a test from one language to another. These would all represent real-world examples of items "jumping around." The last sentence of Mercer's quoted above is a key one. Although it seems reasonable on its face, it does not adequately identify the model to which she is committed. In order to do so, it would have to read, "The difficulty levels of *all items at once* are set by the cultural base from which they are drawn, and these levels are augmented *by a constant amount* for all items when a test is administered to a population from a slightly different cultural base." But this description would begin to reveal the improbable nature of the model of diffusion that her definition of cultural bias requires.

When giving examples of items that purportedly demonstrate the Anglo-centric bias of IQ tests, critics often quote verbal items that are intended to be difficult even for whites. Thus, counsel for plaintiffs in *Larry P.* asked Doppelt (1977:58) about the item, "Who wrote Romeo and Juliet?" and Mercer (1977:1616), during her testimony, quoted the item "Who is Longfellow?" The critics are generally silent, however, about where in the culture the "bits of knowledge" reside that account for performance on the outlandish items that appear in nonverbal IQ tests, and about how it is that cultural bias accounts for the fact that whites show an equal or greater advantage over blacks in this nonverbal content realm that is so esoteric to all cultures.

Non-Anglo cultural groups with IQ means well over 100 are not mentioned. The American Jews in Bachman's (1970:Table E-4-5) representative sample of 10th grade boys show a verbal IQ mean of 112.8, which agrees closely with the earlier run of findings based on less well-defined samples (Nardi, 1948). (This is seen when the mean and standard deviation of Bachman's whites are equated to 101.8 and 16.4, respectively. These values, which are the most representative estimates of the white parameters, come from the 1937 Stanford-Binet normative sample. See Terman and Merrill, 1960:Fig.4) Lynn (1977) has examined the Japanese standardization of the Performance tests, which do not require translation, in the WISC and other Wechsler tests, and concluded that the overall results are consistent with a mean Japanese IQ of 106.6. In view of these facts, according to Mercer's logic in which the mean difference between groups is read directly as evidence of cultural bias, the instruments in question ought to be known instead as "Jewish-Japanese-Anglo" tests. Obviously, this would not have the same rhetorical impact.

After summarizing his work on internal validity, which included examination of test reliabilities for both races, loadings of items on the first principal component, choice of distractors for test items, and the components of item-group interaction mentioned earlier, Jensen (1976:346) concluded with the following statement:

> Claims based on subjective, armchair surmise and speculation about cultural biases in specific test items—the sole method of those critics of tests who wish to foster the myth of cultural bias—are proven false by objective evidence. . . . Culturally loaded—of course. But not culturally biased. The distinction is crucial. The myth of culture bias thrives on obscuring this distinction.

Racial Comparisons in Status Attainment Models Involving IQ and SES

Guterman (1979) has raised the question of the "legitimacy" of using IQ as a variable in stratification research, should the tests be culturally biased. He has not explicated the consequences of such a hypothetical situation, however, and of how they might call the legitimacy of the research into question. The consequences would seem to depend on whether IQ tests

are used for selection at any stage of the attainment process or whether IQ simply describes a process that tests themselves do not influence.

If biased tests are used for selecting personnel, consequences would depend on whether the tests predict adequately within race but not across race or are inadequate in both respects. Under the former possibility, blacks, for example, could be correctly ordered among themselves but underpredicted in comparison to whites, perhaps by a constant amount. In this case, paths involving selection according to test results and later paths not involving explicit selection as well should behave similarly in models for each race. Under the second possibility, with the tests inadequate in both respects, only paths involving explicit selection should behave similarly for both races; blacks would be selected according to their tested IQ, just as whites, whether the scores were valid within race or not. However, once released from the stage of the attainment process in which selection occurs according to test results, blacks should begin to sort themselves out according to their true ability. Presumably, status would be attained at the later points by informal processes that do not take test scores explicitly into account. At these later stages of the attainment sequence, the IQ variable would play merely a descriptive role (for the researcher). Since the true ability of the blacks would not have the same correlation with their test scores as in the case of whites, tested IQ should behave quite differently for the two races in these parts of the model.

If tests are never used for selection, and IQ simply describes a process which the tests themselves do not influence, all relevant paths should be similar for blacks and whites when IQ is equally valid within race, even if blacks are underpredicted by a constant amount. Recognition of test bias in this case would depend on mean levels rather than correlations. But if IQ were not valid even among blacks, all of the relevant paths should differ from those for whites, just as they should in the prior situation when blacks were released from selection determined by score. Racial comparisons of paths in these situations should yield some spectacular differences, such as occasional reversals of sign and large discrepancies in strength of effects.

It is not clear that the concept of "legitimacy" is applicable to any of these situations, inasmuch as the models are intended to describe what happens whatever the reason. The possibility clearly exists that the models might reveal effects that signal the presence of bias in the tests—surely this is no cause for questioning the legitimacy of the presence of IQ in the models. Perhaps Guterman has imported the concept of "legitimacy" into stratification research from considerations having to do with the use of biased tests in selection. However, there seems to be no reason to confuse the research context with the applied context.

The question of the behavior of IQ across class or race in status attainment models raised by Guterman is an intriguing one, however. Because sociological status attainment studies employ a greater variety of predictive criteria than the studies of external validity in the psychological literature reviewed above, it seems appropriate to incorporate such studies into any general discussion of IQ test bias.

Accordingly, we have tried to identify studies that have estimated status attainment models for both blacks and whites, that have included measures of IQ, and that have published data from which standardized partial regression coefficients (betas) could be calculated. In what follows, we compare the direct effects of IQ with the direct effects of SES at various points in the status attainment process, and then compare the outcomes of the IQ/SES comparison for blacks with the outcomes for whites in both the same study and different studies. Standardized rather than unstandardized partial regression coefficients are employed because the fundamental comparison always involves variables for the same population, and standardized coefficients adjust for differences in scale of measurement. This comparison is accomplished by dividing the IQ beta by the SES beta from the same study at each point in the model, yielding an IQ/SES ratio as a measure of the relative importance of the two variables within each race. Using a ratio of regression weights reduces the problem created by different sample variances when stan-

dardized coefficients are compared across samples, because now the effects of the variance are likely to be present in both numerator and denominator of the IQ/SES ratio, where they would tend to neutralize each other. By thus comparing IQ test results with SES as a measure of social background, we obtain a simple descriptive statistic that is sensitive to potential variations across race in the meaningfulness of test scores, as well as one that reveals the strength of the scores vis-à-vis their logical competitor as a background determinant of status attainment. Other comparisons are possible, e.g., of raw regression coefficients, but the one we have chosen compresses more aspects of the process into a single number.[1]

Figure 1 displays two simple path models suitable for examining the effects of IQ and SES. Two models are specified in order to make use of some studies that did not include measures of educational aspirations or expectations. The variables included in the models are strategically chosen ones; they are variables for which IQ is known to be important from previous status attainment research. The direct paths with which the present analysis is concerned are represented by solid lines.

Table 1 brings together the IQ/SES ratios from equations that we have reestimated from seven studies for both path models in Figure 1. Sample sizes indicated in the notes of the table are only approximate because detailed case bases were not always reported. SES is represented by three usual variables or a composite index whenever that was available from the original report. When multiple measures of SES were reported, we constructed a sheaf coefficient (Heise, 1972) to summarize the effects of the individual SES measures and provide a composite index. In our equations, only one SES variable was present in the model at a time, with the composite index counting as one variable for this purpose. Cases in which the SES beta was negative, indicated by asterisks, have been omitted from the table on the grounds that this reflects a peculiarity of the data inconsistent with usual theory concerning the role of SES. Many of such "negative" coefficients, of course, are quite close to zero. Negative betas for IQ, however, have been retained (only four cases occurred). We have accepted beta coefficients regardless of whether or not they met conventional levels of significance. It should also be noted that the table glosses over usual distinctions between variables in the original studies. For example, "Head's education" is "Father's education" in some studies that have a separate measure for "Mother's education."

Each panel of Table 1 represents a separate equation in the models. Panel A reveals the relative effects of IQ and SES on GPA. It should be noted that 10 of the 19 asterisks (negative SES betas) occur in conjunction with a study (Howell and Frese, 1979) that severely restricted the variance of SES in its design. Alternatively, 10 of the 19 asterisks involve parental education for blacks, which could well function poorly as an index of SES for past generations. One could also account for 13 asterisks by potential instability due to small sample sizes, those less than 200.

Overall, the effect of IQ on GPA is substantially greater than that of SES for both races. In no case does SES have a greater effect than IQ. If we exclude IQ/SES ratios greater than or equal to 10.0, on the grounds that IQ is probably not really that much more powerful than SES and hence such results simply reflect unusually small regression weights in the denominators, we can conveniently summarize the remaining entries by averaging them within race, but taking care to include the composite SES index only when other SES variables were not available in the table. There were 14 such entries for whites and four for blacks, yielding practically identical mean ratios for whites and blacks, respectively, of 4.8 and 4.3. According to these results, the relative roles of IQ and socioeconomic background appear to be about the same for both races.

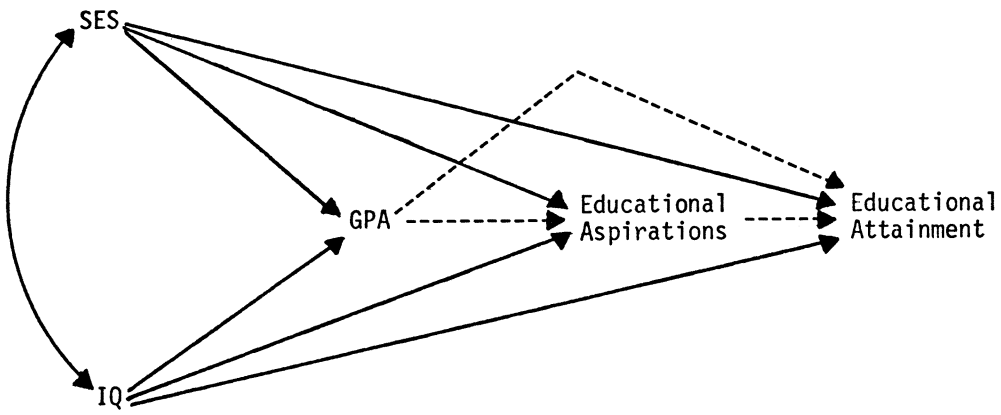
It should be noted at this point that because of the stronger effect of IQ on GPA (see panel A), GPA partials out more of

[1] These calculations ignore differential reliability by race in the reporting of SES. If reports of family SES by blacks contain more measurement error, the black ratio will be biased upward relative to whites, all else being equal. Given the wide range of measures reported, a systematic correction for this problem is not feasible.

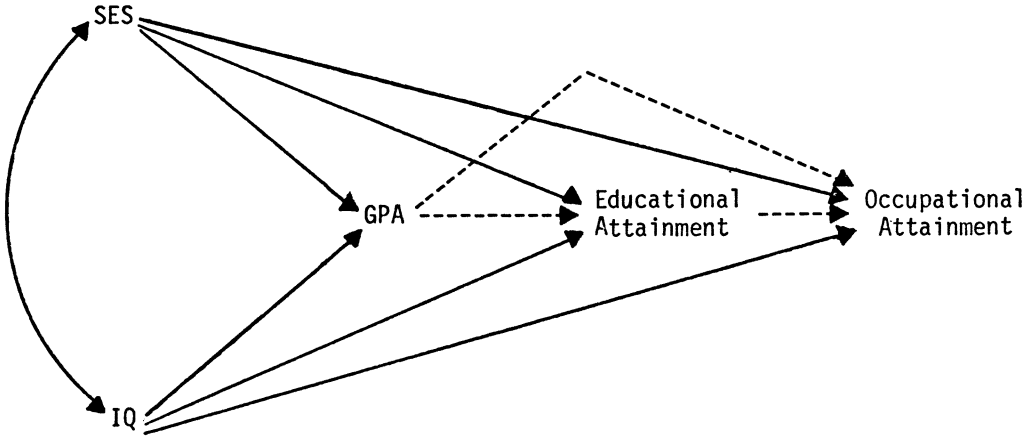
Figure 1

Two Common Models of Status Attainment

Path Model 1



Path Model 2



the effect of IQ than of SES. This has an effect on direct path coefficients later on in the sequence, where GPA is held constant. Later IQ/SES ratios should not be misinterpreted, therefore, as reflecting the relative *total* contributions of IQ and SES, much of which is channeled through the unexamined paths indicated by broken lines in Figure 1. Effects in these unexamined paths are, of course, foreshadowed by earlier direct paths, so there is little to be gained for present purposes by discussing them. Although we could have presented the analysis in terms of total effects, we elected to study direct effects on the grounds that these revealed more subtle aspects of the status attainment process with respect to the relative roles of IQ and

Table 1. The Relative Importance of Direct Paths of IQ and Socioeconomic Status in Determining Indicated Outcome Variable for Blacks and Whites: Standardized Regression Weight for IQ Divided by Standardized Regression Weight for SES

	Composite SES		Father's Occup.		Head's Educ.		Mother's Educ.	
	Whites	Blacks	Whites	Blacks	Whites	Blacks	Whites	Blacks
A. Grade Point Average (Path Models 1 and 2)								
Males:[a]								
Jr. High	1.1	2.0	1.9	5.1	2.0	*	—	—
Sr. High	1.8	1.9	3.4	113.0	3.7	*	—	—
Males[b]	2.9	1.7	—	—	3.4	*	3.6	2.6
Males[c]	19.3	*	—	—	—	—	—	—
Males[d]	—	—	13.2	1.1	—	—	—	—
Males[e]	*	8.3	—	—	—	—	—	—
Males[f]	5.3	7.4	9.2	37.0	5.7	*	6.9	61.0
Females[f]	8.0	*	25.2	23.8	8.8	*	9.8	*
Males[g]	2.6	*	*	*	20.7	*	3.3	*
Females[g]	2.3	*	3.6	*	2.6	*	*	*
B. Educational Aspirations or Expectations (Path Model 1)								
Males[a]								
Jr. High	0.9	0.2	1.1	0.8	1.2	0.6	—	—
Males[b]	0.4	−0.1	—	—	0.5	−0.1	0.6	−0.1
Males[c]	0.4	−1.0	—	—	—	—	—	—
Males[e]	*	3.3	—	—	—	—	—	—
Males[f]	0.4	1.0	0.6	2.9	0.6	1.2	0.7	1.3
Females[f]	0.3	0.5	0.8	1.2	0.5	0.8	0.4	0.6
Males[g]	0.5	0.4	0.6	0.7	1.2	2.3	5.1	0.4
Females[g]	0.8	0.4	0.9	3.4	1.1	1.4	*	0.3
C. Educational Attainment, Model Includes Educational Aspirations/Expectations (Path Model 1)								
Males:[a]								
Jr. High	0.4	2.7	0.5	4.1	0.7	6.1	—	—
Sr. High	0.3	1.0	0.3	1.3	0.5	1.3	—	—
Males[c]	0.9	1.4	—	—	—	—	—	—
D. Educational Attainment, Model Excludes Educational Aspirations/Expectations (Path Model 2)								
Males:[a]								
Jr. High	0.5	2.2	0.7	3.7	0.8	4.7	—	—
Sr. High	0.4	1.0	0.5	1.2	0.6	1.2	—	—
Males[c]	0.8	1.0	—	—	—	—	—	—
Males[d]	—	—	2.4	8.6	—	—	—	—
E. Occupational Attainment, Model Excludes Educational Aspirations/Expectations (Path Model 2)								
Males[d]	—	—	2.6	1.6	—	—	—	—

[a] Kerckhoff and Campbell, 1977a, 324–390 whites, 79–113 blacks. In the table, Jr. High and Sr. high refer to two measures of GPA for the same sample. Senior high school results are omitted from panel B because they are out of the causal sequence indicated by the path model.
[b] Kerckhoff and Campbell, 1977b, 987 whites, 74 blacks.
[c] Portes and Wilson, 1976. 1,957 whites, 256 blacks.
[d] Porter, 1974. 14,891 whites, 435 blacks.
[e] C. Gordon, no date. 1,149 whites, 535 blacks. Somer's d's have been substituted for correlations in estimating regression weights involving GPA.
[f] DeBord, Griffin, and Clark, 1977. 1,014 white males, 439 black males, 1,025 white females, and 550 black females.
[g] Howell and Frese, 1979. 187 white males, 193 black males, 183 white females, and 166 black females. Samples were entirely lower class for both races.
* Indicates negative beta coefficient for SES.

SES. Our ratios in Table 1, therefore, are conservative in terms of showing the potency of IQ, particularly in later stages of the model, compared to what they would have shown had the analysis been conducted in terms of total effects.

Panel B shows that the relative effects of IQ and SES on educational aspirations or expectations are similar for both races. Average ratios, calculated as described above, and based on 16 entries for whites and 18 for blacks, are 1.0 and 1.1, respectively. These include the only negative entries in Table 1, which appeared in columns for blacks. Note that the average relative effect of IQ for blacks exceeds that for whites, even when the negative entries for blacks are counted. Two of these negative entries occur in cells that had negative values for SES in panel A (see asterisks). It is likely that the earlier negative outcome for SES contributes to the later negative outcome for IQ.

Panel C shows that the relative effect of the direct path of IQ on educational attainment, with educational aspirations included, is greater for blacks. Mean ratios are 0.6 for whites and 2.8 for blacks, based on five entries each. With educational aspirations excluded, in panel D, the mean ratio for whites is 1.0 and for blacks 3.4, based on six entries each, with educational attainment as the dependent variable. The two equations represented by panels C and D are the first in which the relative effects become somewhat dissimilar across race, as judged by the means. However, the evidence in both cases shows that IQ relative to SES has greater impact for blacks than whites, and that in the case of blacks, IQ has a stronger direct effect than SES on educational attainment.

Panel E contains just one entry for each race. Both ratios are roughly of the same magnitude, although that for whites is greater. IQ has a greater direct effect on occupational attainment than SES for both whites and blacks.

Inspection of the panels in Table 1 suggests that the variance of results within race across different studies is often greater than the variance across race within studies. This would indicate that when racial differences in IQ/SES ratios are viewed against the general run of differences from study to study, they are not particularly large (an unpresented analysis bears this out). It seems reasonable to conclude that the direct effects of IQ, relative to SES, are common to both races.

CONCLUSION

We have reviewed the major kinds of evidence concerning cultural bias in IQ tests and found no sign that they are in any meaningful sense unfair to blacks or lower class whites. We have also demonstrated that prominent test critics endorse a definition of bias that opportunistically relies on considerations of face validity for purposes of public argument, but more fundamentally on the interpretation of mean differences as evidence that a global diffusion process fails to convey test content to blacks as well as it does to whites. We have argued that studies of internal validity place important limits on the nature of the diffusion process that restrict this process to a quite unrealistic form. Once so restricted, the present state of diffusion of test content is tantamount to a description of the societal distribution of g.

Had test score differences between groups been simply a reflection of test sophistication as discussed, for example, by Vernon (1947:61) and by the Scottish Council for Research on Education (1958:39–47), it would have implied that the differences could be disregarded in practical application, particularly those in which individuals are selected for exposure for the first time to relatively new learning situations. An analogous result was reported by the Scottish Council (1949:136), which attributed verbal group test gains between its 1932 and 1947 mental surveys to test sophistication, noting that "the increase in the verbal score has no counterpart in Binet IQ." Apparently, the lower–scoring 1932 cohort was not really at a relative disadvantage in terms of true intelligence. Unfortunately, the problem of social class and race differences in test scores is not going to be as simple as that, for as Jensen (1973) puts it, "The differences are real."

We have also investigated important

stages of the status attainment process, and found no indication that IQ is relatively less important for blacks than whites in comparison to socioeconomic background. Quite the contrary, according to our conservative analysis IQ tends to be as important as, or more important than, SES throughout the models for both races, but especially for blacks. Simple summary statistics indicated that there was little difference between the races in the relative importance of the two variables in direct paths. In particular, variation between races in the IQ/SES ratio turned out to be more modest than variation between studies. These findings extend the studies of external validity to new criteria of special importance to sociologists. In concurrence with previous studies, no indications of racial bias in IQ tests were found. This evidence could be nullified only by assuming that explicit selection on IQ scores occurs throughout our status attainment models, as discussed earlier. Since the use of IQ tests for selection purposes is typically a matter of public record, such an assumption seems highly unrealistic.

The ultimate significance of the failure to find bias in IQ tests will depend on how easily changed IQ eventually proves to be, on a proper understanding of the magnitude of group differences, on assessments of the importance of IQ in determining other outcomes, and on the social importance of the outcomes so determined.

REFERENCES

Ammons, R. B., and C. H. Ammons
1962 "The Quick Test (QT): Provisional manual." Psychological Reports 11:111–161.

Angoff, William H., and Susan F. Ford
1973 "Item-race interaction on a test of scholastic aptitude." Journal of Educational Measurement 10:95–105.

APA Task Force on Employment Testing of Minority Groups
1969 "Job testing and the disadvantaged." American Psychologist 24:637–650.

Bachman, Jerald G.
1970 Youth in Transition, Vol. 2: The Impact of Family Background and Intelligence on Tenth-Grade Boys. Ann Arbor: Institute for Social Research, University of Michigan.

Bhushan, Vidya
1974 "Adaptation of an intelligence test from English to French." Journal of Educational Measurement 11:43–48.

Blum, Milton L.
1953 "General Aptitude Test Battery." Pp. 685–690 in O.K. Buros (ed.), The Fourth Mental Measurements Yearbook. Highland Park, N.J.: Gryphon Press.

Boehm, Virginia
1977 "Differential prediction: A methodological artifact?" Journal of Applied Psychology 62:146–154.

Boney, J. Don
1966 "Predicting the academic achievement of secondary school Negro students." Personnel and Guidance Journal 44:700–703.

Burt, Cyril.
1970 "The genetics of intelligence." Pp. 15–28 in W. B. Dockrell (ed.), On Intelligence. London: Methuen.

Campbell, Joel T., Ronald L. Flaugher, Lewis W. Pike, and Donald A. Rock
1969 "Bias in selection tests and criteria studied by ETS and U.S. Civil Service." ETS Developments 17:2.

CBS
1975 "The IQ Myth." CBS Reports, April 22, 1975. Transcript.

Chase, Clinton I., and Richard C. Pugh
1971 "Social class and performance on an intelligence test." Journal of Educational Measurement 8:197–202.

Cleary, T. Anne
1968 "Test bias: Prediction of grades of Negro and white students in integrated colleges." Journal of Educational Measurement 5:115–124.

Cleary, T. Anne, and Thomas L. Hilton
1968 "An investigation of item bias." Educational and Psychological Measurement 28:61–75.

Cleary, T. Anne, Lloyd G. Humphreys, S. A. Kendrick, and Alexander Wesman
1975 "Educational uses of tests with disadvantaged students." American Psychologist 30:15–41.

Coleman, James S., Ernest Q. Campbell, Carol J. Hobson, James McPartland, Alexander M. Mood, Frederic D. Weinfeld, and Robert L. York
1966 Equality of Educational Opportunity. Washington, D.C.: U.S. Government Printing Office.

Dalton, Starrette
1974 "Predictive validity of high school rank and SAT scores for minority students." Educational and Psychological Measurement 34:367–370.

DeBord, Larry W., Larry Griffin, and Melissa Clark
1977 "Race and sex influences in the schooling process of rural and small town youth." Sociology of Education 50:85–102.

Doppelt, Jerome
1977 "Deposition." Reporters' Transcript, Larry P. et al. vs. Wilson Riles et al., United States District Court, Northern District of California.

Dvorak, Beatrice J.
1956 "The General Aptitude Test Battery." Personnel and Guidance Journal 35:145–152.
Gael, Sidney and Donald L. Grant
1972 "Employment test validation for minority and nonminority telephone company service representatives." Journal of Applied Psychology 56:135–139.
Goldman, Roy D., and Barbara Newlin Hewitt
1976 "Predicting the success of black, Chicano, oriental, and white college students." Journal of Educational Measurement 13:107–117.
Goldman, Roy D., and Regina Richards
1974 "The SAT prediction of grades for Mexican-American versus Anglo-American students at the University of California, Riverside." Journal of Educational Measurement 11:129–135.
Gordon, Chad
No date Looking Ahead: Self-Conceptions, Race and Family as Determinants of Adolescent Orientation to Achievement. Arnold M. and Caroline Rose Monograph Series, American Sociological Association.
Gordon, Mary Agnes
1953 "A study in the applicability of the same minimum qualifying scores for technical schools to white males, WAF, and Negro males." Technical Report 53–34, Human Resources Research Center, Lackland Air Force Base, San Antonio, Texas.
Gordon, Robert A.
1975 "Examining labeling theory: The case of mental retardation." Pp. 83–146 in W. R. Gove (ed.), The Labelling of Deviance: Evaluating a Perspective. Beverly Hills, Calif.: Sage/Halstead.
1976 "Prevalence: The rare datum in delinquency measurement and its implications for the theory of delinquency." Pp. 201–284 in M. W. Klien (ed.). The Juvenile Justice System. Beverly Hills, Calif.: Sage.
Grant, Donald L. and Douglas W. Bray
1970 "Validation of employment tests for telephone company installation and repair occupations." Journal of Applied Psychology 54:7–14.
Guterman, Stanley S.
1979 "I.Q. tests in research on social stratification: The cross-class validity of the tests." Sociology of Education 52:163–173.
Heise, David R.
1972 "Employing nominal variables, induced variables, and block variables in path analyses." Sociological Methods and Research 1:147–173.
Hills, John R. and Julian C. Stanley
1968 "Prediction of freshman grades from SAT and from level 4 of SCAT in three predominantly Negro state colleges." Proceedings, 76th Annual Convention, American Psychological Association:241–242.
1970 "Easier test improves prediction of black students' college grades." Journal of Negro Education 39:320–324.
Howell, Frank M., and Wolfgang Frese
1979 "Race, sex, and aspirations: Evidence for the 'race convergence' hypothesis." Sociology of Education 52:34–46.
Humphreys, Lloyd G.
1973 "Statistical definitions of test validity for minority groups." Journal of Applied Psychology 58:1–4.
1975 "Addendum." American Psychologist 30:95–96.
Hunter, John E., Frank L. Schmidt, and Ronda Hunter
Forthcoming "Differential validity of employment tests by race: A comprehensive review and analysis." Psychological Bulletin.
Jensen, Arthur R.
1969 "How much can we boost IQ and scholastic achievement?" Harvard Educational Review 39:1–123.
1973 "The differences are real" Psychology Today 7 (December):80–86.
1974 "How biased are culture-loaded tests?" Genetic Psychology Monographs 90:185–244.
1976 "Test bias and construct validity." Phi Delta Kappan 58:340–346.
1977 "An examination of culture bias in the Wonderlic Personnel Test." Intelligence 1:51–64.
Kallingal, Anthony
1971 "The prediction of grades for black and white students at Michigan State University." Journal of Educational Measurement 8:263–265.
Kamin, Leon J.
1974 The Science and Politics of I.Q. Potomac, Md.: Erlbaum/Halsted.
1977 "Testimony." Reporters' Daily Transcript, Larry P. et al. vs. Wilson Riles et al., United States District Court, Northern District of California.
Katzell, Raymond A., and Frank J. Dyer
1977 "Differential validity revived." Journal of Applied Psychology 62:137–145.
Kerckhoff, Alan C., and Richard T. Campbell
1977a "Black-white differences in the educational attainment process." Sociology of Education 50:15–27.
1977b "Race and social status differences in the explanation of educational ambition." Social Forces 55:701–714.
Lerner, Barbara
1976 "Washington v. Davis: Quantity, quality and equality in employment testing." Supreme Court Review 1976:263–316.
Linn, Robert L.
1975 "Test bias and the prediction of grades in law school." Journal of Legal Education 27:293–323.
Lynn, Richard
1977 "The intelligence of the Japanese." Bulletin of the British Psychological Society 30:69–72.
Maier, Milton, and Edmund F. Fuchs
1975 "Effectiveness of selection and classification testing." U.S. Army Research Institute

for the Behavioral and Social Sciences Research Report, 1973, No. 1179. Catalogue of Selected Documents in Psychology 5:209.

McNemar, Quinn
1942 The Revision of the Stanford–Binet Scale. Boston: Houghton Mifflin.

Medley, Donald M., and Thomas J. Quirk
1974 "The application of a factorial design to the study of cultural bias in general culture items on the National Teacher Examination." Journal of Educational Measurement 11:235–245.
1978 "Testimony." Reporters' Daily Transcript, Larry P. et al. vs. Wilson Riles et al., United States District Court, Northern District of California.

Mercer, Jane R.
1973 Labeling the Retarded. Berkeley: University of California Press.

Mercer, Jane R., and June F. Lewis
Forthcoming SOMPA: System of Multi–Cultural Pluralistic Assessment. Riverside, Calif.: Institute for Pluralistic Assessment Research and Training.

Miele, Frank
1979 "Cultural bias in the WISC." Intelligence 3:149–164.

Nardi, Noah
1948 "Studies in intelligence of Jewish children." Jewish Education 19:41–51.

Nichols, Paul Leslie
1970 "The effects of heredity and environment on intelligence test performance in 4 and 7 year old white and Negro sibling pairs." Unpublished Ph.D. dissertation, University of Minnesota.

O'Connor, Edward J., Kenneth N. Wexley, and Ralph A. Alexander
1975 "Single–group validity: Fact or fallacy?" Journal of Applied Psychology 60:352–355.

Pfeifer, C. Michael, Jr. and William E. Sedlacek
1971 "The validity of academic predictors for black and white students at a predominantly white university." Journal of Educational Measurement 8:253–261.

Porter, James N.
1974 "Race, socialization and mobility in educational and early occupational attainment." American Sociological Review 39:303–316.

Portes, Alejandro, and Kenneth L. Wilson
1976 "Black–white differences in educational attainment." American Sociological Review 41:414–431.

Schmidt, Frank L., John G. Berner, and John E. Hunter.
1973 "Racial differences in validity of employment tests: Reality or illusion?" Journal of Applied Psychology 58:5–9.

Scottish Council for Research in Education
1949 The Trend of Scottish Intelligence: A Comparison of the 1947 and 1932 Surveys of the Intelligence of Eleven–Year–Old Pupils. London: University of London Press.
1958 Educational and Other Aspects of the 1947 Scottish Survey. London: University of London Press.

Stanley, Julian C.
1969 "Plotting ANOVA interactions for ease of visual inspection." Educational and Psychological Measurement 29:793–797.
1971a "Predicting college success of the educationally disadvantaged." Science 171 (19 February):640–647.
1971b "Predicting college success of educationally disadvantaged students." Pp. 58–77 in Stephen J. Wright (ed.), Barriers to Higher Education. New York: College Entrance Examination Board.

Stanley, Julian C. and Andrew C. Porter
1967 "Correlation of scholastic aptitude test score with college grades for Negroes versus whites." Journal of Educational Measurement 4:199–218.

Temp, George C.
1971 "Validity of the SAT for blacks and whites in thirteen integrated institutions." Journal of Educational Measurement 8:245–251.

Tenopyr, M. L.
1967 "Race and socioeconomic status as moderators in predicting machine-shop training success." Paper presented at the meeting of the American Psychological Association, Washington, D.C., September.

Terman, Lewis M. and Maud A. Merrill
1937 Measuring Intelligence. Boston: Houghton Mifflin.
1960 Stanford–Binet Intelligence Scale: Manual for the Third Revision Form L–M. Boston: Houghton Mifflin.

Thomas, Charles L.
1971 "The relative effectiveness of high school grades and standardized test scores for predicting college grades of black students." Unpublished Ph.D. dissertation, Department of Education, Johns Hopkins University.

Vernon, P. E.
1947 "The variations of intelligence with occupation, age, and locality." British Journal of Statistical Psychology 1:52–63.

Williams, Robert L.
1972 "Black intelligence test of cultural homogeneity." Unpublished manuscript. St. Louis, Missouri.