DOCUMENT RESUME

ED 209 252 TM 810 684

AUTHOR Reynolds, Cecil R.
TITLE Test Bias: In God We Trust, All Others Must Have Data.
PUB DATE Aug 81
NOTE 53p.; Paper presented at the Annual Meeting of the American Psychological Association (Los Angeles, CA, August, 1981).

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Elementary Secondary Education; *Intelligence Tests; Personality Measures; *Test Bias; *Testing Problems; Test Validity
IDENTIFIERS Wechsler Intelligence Scale for Children (Revised)

ABSTRACT

The cultural test bias hypothesis represents the contention that all ethnic or racial group differences on mental tests are due to inherent, artifactual biases produced within the tests through flawed psychometric methodology. This address focuses on an empirical evaluation of the cultural test bias hypothesis, especially emphasizing the construct validity and criterion-related validity of these tests. A considerable body of literature exists failing to substantiate cultural bias against native-born American ethnic minorities with regard to the use of well-constructed, adequately standardized intelligence and aptitude tests. With respect to personality scales, the evidence is far more preliminary and thus considerably less conclusive. Despite this evidence, psychologists need to keep abreast of new findings in the area. As new techniques and better methodology are developed and more specific populations examined, the now seen as random, infrequent findings of bias may become better understood and seen to display a correctable pattern. (Author/BW)

TEST BIAS: IN GOD WE TRUST, ALL OTHERS MUST HAVE DATA

Cecil R. Reynolds, Ph.D.
Department of Educational Psychology
704 Harrington Education Center
Texas A & M University
College Station, Texas 77843

Invited address for the APA Division of Evaluation and Measurement to the annual meeting of the American Psychological Association, Los Angeles, August, 1981.

The cultural test bias hypothesis represents the contention that all ethnic or racial group differences on mental tests are due to inherent, artifactual biases produced within the tests through flawed psychometric methodology. Group differences are believed then to stem from characteristics of the tests and to be totally unrelated to any actual differences in the psychological trait in question. The resolution or evaluation of the validity of the cultural test bias hypothesis is one of the most crucial scientific questions facing psychology today.

Bias in mental tests has many implications for individuals including pupil misassignment to educational programs, unfair denial of admission to college, graduate, and professional degree programs, and the inappropriate denial of employment. The scientific implications are even more substantive. There would be dramatic implications for psychological research and theory if the cultural test bias hypothesis is correct: The principal research of the past 100 years in the psychology of human differences would have to be dismissed as confounded and largely artifactual since much of the work is based on standard psychometric theory and testing technology (Reynolds, 1980a; Reynolds & Brown, in press). This would in turn create major upheavals in applied psychology, since the foundations of clinical, counseling, industrial, and school psychology are all strongly tied to the basic academic field of individual differences.

This crucial issue must be resolved, and resolved on the basis of empirical study. This is not an easy task, for in the

area of test bias, it is easier "... to convey wrong impressions, easier to fan hot coals of bigotry than to turn on lights of understanding" (Horn & Goldsmith, 1981, p.322). Increasingly the issue of cultural test bias has become an emotional, political, and legal one. Writers on both sides of the issue have been given to emotional tirades. Williams (1974) has referred to those who use psychological tests with black children as "white pimp and hustler type psychologists" whose intent is the abuse and dehumanization of black people. On the other hand, many of the early, well-known psychometricians (e.g., Yerkes, Terman, and Pearson) held openly racist positions and made what are now generally seen as inappropriate interpretations of race differences on mental tests in support of restrictive immigration laws and other political tragedies. Over the past several decades, the issue of bias has moved into legislatures and the courts. New York State has enacted "truth-in-testing" legislation and similar bills are being considered at the Federal level. Several major court decisions have also recently been handed down attempting to resolve the question of bias (Larry P., 1979; PASE, 1980). Emotional, political, and legal attempts to resolve the validity of the cultural test bias hypothesis are inherently unacceptable.

Take for example the legal response to the question "Are intelligence tests used to diagnose mental retardation biased against cultural and ethnic minorities?" In California in 1979 (Larry P., 1979) the answer was "Yes" but in Illinois in 1980 (PASE, 1980) the response was "No." Thus two federal district courts of equivalent standing have heard nearly identical

cases, with many of the same witnesses espousing much the same testimony, and reached the exact opposite conclusions. Much of the emotionality and subjective influence of committment to doctrine was evident in the testimony of the plaintiffs' witnesses, so much so that Judge Grady (PASE, 1980) essentially dismissed most of the "expert" witness testimony as irrelevant, incompetent, or biased in its own right. It should come as no surprise that Grady dismissed much of the testimony after reviewing its contents.

One witness for the plaintiffs objected to asking black children in what direction the sun sets because so many of these children live in high rise tenements and may never have been on the west side of their building. Another well-known psychologist testifying for the plaintiffs objected to the use of a picture of an ordinary hair comb explaining that black children would not recognize the item as a comb since they are only exposed to "Afro-type" combs. (One should also recall that this item is the easiest on the subtest where it appears and that the test is designed for children aged 6 yrs. to 16½ yrs.). The famous "fight" item (Item 6 on the WISC-R) was again criticized despite considerable evidence that this item is easier or at least no more difficult for black than white children (e.g., see Mercer, in press; Reynolds, 1982; Sandoval, Zimmerman, & Woo-Sam, 1980).

Though current opinion on the cultural test bias hypothesis is quite divergent, ranging from those who consider it to be for the most part unresearchable (e.g., Hirsch, 1981; Schoenfeld, 1974)

to those who consider the issue settled (e.g., Jensen, 1980), it seems clear that empirical analysis of the hypothesis must be undertaken. However difficult full objectivity may be in science, we must make every attempt to view all socially, politically, and emotionally charged issues from the perspective of rational scientific inquiry. We must also be prepared to accept scientifically valid findings as real, whether we like them or not. Otherwise we certainly risk psychology becoming an impotent field whose issues are not resolved by scholars in the courts of scientific inquiry but by judges in courts of law, and whose practitioners opinions are of no more validity than the faith of devotees to a socio-political doctrinaire.

A number of factors need to be considered when evaluating the cultural test bias hypothesis, not the least of which is the historical perspective from which many concerned groups must view the issues. This and other issues must be considered but cannot be presented within the scope of this address. The interested reader will need to pursue several sources to achieve a balanced view (e.g., Hirsch, 1981; Jensen, 1980a,b; Reynolds, 1982; Reynolds & Brown, in press a.b). The present address will focus on an empirical evaluation of the cultural test bias hypothesis that reflects the research program in this area of myself and my colleagues, though the work of others will be referred to when necessary. Prior to proceeding however, I would like to point out a paradox in the current literature on bias regarding bias in intelligence and in personality measures.

CONTRADICTORY CLAIMS OF TEST BIAS

The criticisms of intelligence and other aptitude tests are

well-known and reviewed in a variety of sources. Recently (Reynolds, 1982), I collapsed these various criticisms into 6 basic groups.

1) The content of the tests is unfamiliar to and inappropriate for use with minority children.
2) The standardization samples of the tests include minorities in insufficient numbers for them to significantly impact item selection.
3) Examiner and language bias is present since most psychologists are white and speak only standard English, intimidating and confusing minority children.
4) Inequitable social consequences result when minorities, allready at an economic and educational disadvantage, are relegated to inferior programs because of test performance.
5) The tests measure different attributes when used with children outside of the mainstream, white, middle-class culture.
6) The tests do not predict any important outcomes or future behaviors for minority children.

Psychologists are thus directed to interpret intelligence test performance differently depending upon the race or ethnic background of the child in question. Thus, race or ethnic status takes on the status of a moderator variable. Apparently psychologists have been listening to the hue and cry of bias and do alter their interpretations of tests for blacks and whites.

Reynolds (1982) reviews a number of studies from the cognitive realm indicating psychologists alter recommendations to

special class placement based on race of the child. When faced with obtained IQs, practicing clinicians tend to overestimate the "true IQ" of blacks relative to whites. Further, when IQ and achievement in the classroom are held constant, black children are less likely to be recommended for special class placement than their white counterparts. This particular type of bias works consistently to keep blacks and other minorities out of treatment programs whether the treatment programs are viewed as desirable or undesirable.

Other areas of bias are also now being addressed, one of the most important being potential bias in the evaluation of psychopathology. The potential for cultural bias in personality measures, both objective and projective techniques, has not yet, however received nearly the attention that have cognitive tests. Personality and overt behavior are almost certain to be more culturally determined than are one's intellectual skills. Cross-racial studies of personality scales typically have not been cast into the paradigm of bias but have rather looked at differential responding as reflecting differences in a given personality dimension. Whether the same dimensions of personality and overt behavior exist cross-racially has been little researched. Evidence is now being brought to bear on these issues and will be briefly introduced later. The methodology of research on bias for cognitive measures is equally applicable to standardized personality measures; the question of cultural bias in personality assessment is in dire need of investigation and need not await any further methodological refinements, though

some paradigmatic shifts in thinking may need to occur.

Several interesting studies of bias in the diagnosis and evaluation of psychopathology and behavior have recently appeared (though they did not examine the specific tests in use) that serve well to point out conflicting claims of those who decry the use of tests with minorities. Lewis, Balla, and Shanok (1979) recently reported that when black adolescents are seen in community mental health settings, behaviors symptomatic of schizophrenia, paranoia, and a variety of psychoneurotic disorders are frequently dismissed as only "cultural aberrations" appropriate to coping with the frustrations created by the antagonistic white culture. Lewis et al. further noted that white adolescents exhibiting similar behaviors were given psychiatric diagnoses and referred for therapy and/or residential placements that were not provided blacks. Lewis et al. contend that this failure to diagnose mental illness in the black population acts as bias in the denial of appropriate services. Another study (Lewis, Shanok, Cohen, Kligfeld, & Frisone, 1980) found that "... many seriously psychiatrically disturbed, aggressive black adolescents are being channeled to correction facilities while their equally aggressive white counterparts are directed toward psychiatric treatment facilities" (Lewis et al., 1980, p. 1216). The expressed "failure" of mental health workers to diagnose these black adolescents as emotionally disturbed may be attributed to the critics of psychological testing of minorities. These workers have been told repeatedly that behaviors that are unacceptable in the society-at-large are not only acceptable

in the black culture but adaptive and in some cases necessary.

Plaintiffs' witnesses in Larry P. (1979) and PASE (1980) indicated, for example. that, although it might be appropriate for a white middle class child to respond to a much smaller child who starts a fight by not fighting and by seeking other solutions, black children must respond by fighting back because any other response would be nearly suicidal in the black ghetto culture. Through such criticisms psychologists are led (a) to believe that aggression and violence are not pathological among certain groups and (b) to interpret their behavior and personality test scores differently.

Test interpretations should not be modified on the basis of externally-perceived desirability of programming for one or another group. How can tests be considered biased in the case of regular vs. special education placement and not biased in the case of incarceration vs. mental health treatment? Modifications in test score interpretation cannot ethically be undertaken on the basis of anecdotal or "expert witness" testimony. The decision to modify test score interpretations must ultimately be guided by empirical data. Much has been done in the cognitive arena, but bias in noncognitive measures is a recent consideration. Let us turn now to an examination of the empirical evidence, especially as it pertains to the construct and criterion-related validity of these tests.

BIAS IN CONSTRUCT VALIDITY OF INTELLIGENCE TESTS

There is no single method for the accurate determination of the construct validity of educational and psychological tests.

The defining of bias in construct validity then requires a general statement that can be researched from a variety of viewpoints with a broad range of methodology. The following rather parsimonious definition is proffered:

> Bias exists in regard to construct validity when a test is shown to measure different hypothetical traits (psychological constructs) for one group than another or to measure the same trait but with differing degrees of accuracy.

As is befitting the concept of construct validity, many different methods have been employed to examine existing psychological tests and batteries of tests for potential bias in construct validity. One of the more popular and necessary empirical approaches to investigating construct validity is factor analysis. Hilliard (1979), one of the more vocal critics of IQ tests on the basis of cultural bias, has pointed out one of the potential areas of bias dealing with the comparison of the factor analytic results of test studies across race. "If the IQ test is a valid and reliable test of 'innate' ability or abilities, then the factors which emerge on a given test should be the same from one population to another, since 'intelligence' is asserted to be a set of mental processes. Therefore, while the configuration of scores of a particular group on the factor profile would be expected to differ, logic would dictate that the factors themselves would remain the same" (Hilliard, 1979, p. 53). While certainly not agreeing that identical factor analytic results for an instrument indicate innateness of the abilities being measured,

consistent factor analytic results across populations do provide strong evidence that whatever is being measured by the instrument is being measured in the same manner and is in fact the same construct within each group. The information derived from comparative factor analysis across populations is directly relevant to the use of educational and psychological tests in diagnosis and other decision-making functions. Psychologists, in order to make consistent interpretations of test score data, must be certain that the test(s) measures the same variable across populations.

We have, along with other researchers, undertaken comparative factor analyses of common intelligence tests for blacks and whites, with some studies including Mexican-American and Native American Indians. The most frequently used individual measure of intelligence for school aged children is unquestionably the WISC-R (Wechsler, 1974), so it is appropriate that most research has focussed on this scale and its predecessor, the 1949 WISC. Some of the earliest work in this regard, for the WISC-R, was published by Reschly (1978).

Using a large, random sample, Reschly (1978) compared the factor structure of the WISC-R across four racially identifiable groups: Whites, Blacks, Mexican-Americans, and Native American Papagos, all from the southwestern United States. Consistent with the findings of previous researchers with the 1949 WISC (Lindsey, 1967; Silverstein, 1973), Reschly (1978) reported substantial congruency of factors across race when the two-factor solutions were compared (the two-factor solution typically

delineates Wechsler's a priori grouping of the subtests into a Verbal and a Performance, or nonverbal, scale). The 12 coefficients of congruence for comparisons of the two-factor solution across all combinations of racial groupings ranged only from .97 to .99, denoting factorial equivalence of this solution across groups. Reschly compared three-factor solutions also (three-factor solutions typically relinquish Verbal Comprehension, Perceptual Organization, and Freedom from Distractibility factors), finding congruence only between Whites and Mexican-Americans. These findings are also consistent with previous research with the WISC (Semler & Iscoe, 1966). The "g" factor (representing general intelligence) present in the WISC-R was shown to be congruent across race, as it was also demonstrated by Miele (1979). Reschly concluded that the usual interpretation of the WISC-R Full Scale IQ as a measure of overall, general intellectual ability appears to be equally appropriate for Whites, Blacks, Mexican-Americans, and Native American Papagos. Reschly also concluded that the Verbal/Performance Scale distinction on the WISC-R is equally appropriate across race and that there is strong evidence for having confidence in the integrity of the construct validity of the WISC-R for a variety of populations.

Support for Reschly's (1978) conclusions is available from a variety of other factorial studies of the WISC and WISC-R using many methods of factor analysis. Applying a hierarchical factor analytic method, Vance and Wallbrown (1978) factor analyzed the intercorrelation matrix of the WISC-R subtests for 150 Blacks from the Appalachian region of the U.S., who had been

referred to a psychoeducational clinic. The two-factor heirarchical solution determined for Vance and Wallbrown's (1978) Blacks was highly similar to heirarchical factor solutions determined for the standardization sample of the WISC-R (Wallbrown, Blaha, Wallbrown, & Engin, 1975), the 1949 WISC (Blaha, Wallbrown,& Wherry, 1974), and other Wechsler Scales.

Several more recent studies comparing the WISC-R factor structure across race for normal and referral populations of children have also provided increased support for the generality of Reschly's (1978) conclusions and the results of the other investigators cited above. Oakland and Feigenbaum (1979) factor analyzed the 12 WISC-R subtests' intercorrelations separately for stratified (race, age, sex, SES) random samples of normal White, Black, and Mexican-American children from a large urban school district of the southwestern U. S. Pearson r's were calculated between corresponding factors for each group. For the "g" factor, the Black-White correlation between factor loadings was .95, the Mexican-American-White correlation was .97, and the Black-Mexican American correlation was .96. Similar comparisons across all WISC-R variables produced correlations ranging only from .94 to .99. Oakland and Feigenbaum concluded that the results of their factor analyses "...do not reflect bias with respect to construct validity for these three racial-ethnic...groups" (p. 973).

Gutkin and Reynolds (1981) determined the factorial similarity of the WISC-R for groups of Black and White children from the WISC-R standardization sample. This study is particularly important to

to examine in determining the construct validity of the WISC-R across race due to the sample employed in the investigation. The sample considered of 1868 White and 305 Black children obtained in a stratified random sampling procedure designed to mimic the 1970 U. S. census data on the basis of age, sex, race, SES, geographic region of residence in the U. S., and urban vs. rural residence. Similarity of the WISC-R factor structure across race was investigated by comparing each of the following for the Black and White groups for two-and three-factor solutions: (a) the magnitude of unique variances, (b) the pattern of subtest loadings on each factor, (c) the portion of total variance accounted for by common factor variance, and (d) the percentage of common factor variance accounted for by each factor. Coefficients of congruence comparing the unique variances, the "g" factor, the two-factor, and the three-factor solutions across race all achieved a value of .99. The portion of total variance accounted for by common factor variance varied negligibly for Blacks and Whites being 53% and 51% respectively. The percentage of common factor variance accounted for by each factor in both the two-and three-factor solutions was also strikingly similar across these two racial groups. Gutkin and Reynolds (1981) concluded that for White and Black children, the WISC-R factor structure was essentially invariant and that no evidence of single-group or differential construct validity could be found.

As Table 1 amply demonstrates, this conclusion is not dependent on the particular method of factor comparison employed.

Insert Table 1 About Here

Under all of the six techniques of factor comparison shown in Table 1, conclusions of factorial similarity would have been reached regarding all 3 WISC-R factors for blacks and for whites. These techniques of factor comparison are by far the most frequent to be employed in studies of test bias, and the considerable degree of similarity of outcome across methods should create greater confidence in the findings of researchers using divergent methods of comparison (Reynolds & Harding, 1981).

Other studies are also available comparing the factor structure of the WISC-R across race. Gutkin and Reynolds (1980) and Dean (1979) have also reported strong support for equivalent construct validity of the WISC-R across racial groupings. These writers have reported consistently large coefficients of congruence for 2-factor and 3-factor solutions, "g" factors, and similarity of the strength of these factors across groups.

To provide a summary and to further explore the consistency of the results of cross-race WISC-R factor analyses, Table 2 was developed. To develop Table 2, 2- and 3-factor principal

Insert Table 2 About Here

factor solutions for the WISC-R were derived based only on the scores of the White children in the WISC-R standardization sample, using R^2 as initial communality estimates, and the factors extracted

rotated to a Varimax (orthogonal) solution. The analysis was performed for the 10 regular subtests, 11 subtests excluding Mazes, and the 12 total tests. The first, unrotated principal factor was taken as an estimate of "g." Coefficients of congruence were calculated between corresponding factors for Whites from the standardization sample and corresponding factors based on samples of Blacks and Mexican-Americans from the various studies cited above. Table 2 represents a widely varied set of populations, the methods of initial factor extraction and subsequent rotation are many, and variance of scores used to determine the many factor solutions almost certainly unequal; the comparisons contain an abundant number of the flaws that Mulaik (1972) points out as reasons for failing to find factorial similarity. Even under such adverse statistical conditions, Table 2 clearly demonstrates factorial invariance of the WISC-R across race, thus meeting Hilliard's (1979) criterion of consistent construct validity across race. With regard to school aged children, comparative factor analysis clearly produces no evidence of test bias. Such findings are not exclusive to the WISC-R, though it has been the featured battery thus far.

Results of studies of preschool-aged children yield very similar results with a variety of tests and test batteries (e.g., see Kaufman & DiCuio, 1975; Kaufman & Hollenbeck, 1974; Merz, 1970; Reynolds, 1978, 1979, 1980b). DeFries, Vandenberg, McClearn, Kuse, Wilson, Ashton, and Johnson (1974), based on a factor analysis of 15 mental tests, concluded"...that the structure of

intelligence is also similar for Japanese-Americans and Chinese-Americans."

Other evidence of consistent construct validity of aptitude tests across race has also been recently provided. The definition of bias in construct validity proffered above requires that the accuracy of measurement be constant across groups. Many studies exist showing a high degree of consistency among estimates of internal reliability of these tests across race for blacks, whites, and Mexican-Americans (e.g., Dean, 1977; Jensen, 1977, 1974; Oakland & Feigenbaum, 1979; Sandoval, 1979; Reynolds & Piersel, 1981), though the proper statistical comparison of reliability coefficients is not frequently undertaken (Reynolds, in press). For children, the correlations between age and raw scores is also relatively constant across race (Jensen, 1980; Reynolds, 1980c). Other measures of differential construct have been employed as well and are reviewed in several sources (e.g., Jensen, 1980; Reynolds, 1982; Reynolds & Brown, in press b).

Construct validity of a large number of popular intelligence tests has been investigated across race and sex with a variety of populations of minority and White children and with a divergent set of methodologies. All roads have led to Rome. No consistent evidence of bias in construct validity has been found with any of the many tests investigated. This leads to the conclusion that, for now, the evidence indicates that psychological tests, especially aptitude tests, function in essentially the same manner across

race and sex, the test materials are perceived and reacted to in a similar manner, and that the tests are measuring the same construct with equivalent accuracy for Blacks, Whites, Mexican-Americans, and other native born American ethnic minorities for both sexes. Single group and differential validity have not been found and likely are not an existing phenomenon with regard to well constructed standardized psycholgical and educational tests. This means that test score differences across race are most likely real and not an artifact of test bias. These differences cannot be ignored and, as Miele (1979) has succintly stated, "If this...difference [in test scores] is the result of genetic factors, acceptance of the cultural bias hypothesis would be unfortunate. If the difference is the result of environmental factors, such acceptance would be tragic" (p. 162).

BIAS IN CRITERION RELATED VALIDITY OF INTELLIGENCE TESTS

Evaluating bias in predictive validity of educational and psychological tests is less related to the evaluation of group mental test score differences than to the evaluation of individual test scores in a more absolute sense. This is especially true for aptitude (as opposed to diagnostic) tests where the primary purpose of administration is the prediction of some specific future outcome or behavior. Internal analyses of bias (such as with content and construct validity) are less confounded than analyses of bias in predictive validity, however, due to the potential problems of bias in the criterion measure. Predictive validity

is also strongly influenced by the reliability of criterion measures, which frequently is poor.

Arriving at a consensual definition of bias in predictive validity is also a difficult task. Yet, from the standpoint of the practical applications of aptitude and intelligence tests, predictive validity is the most crucial form of validity in relation to test bias. Much of the discussion in professional journals concerning bias in predictive validity has centered around models of selection. These issues have been debated extensively and need not distract us here. Since the present section is concerned with bias in respect to the test itself and not the social or political justifications of any one particular selection model, the Cleary, Humphreys, Kendrick, and Wesman (1975) definition, with only slight restatement, provides a clear direct statement of test bias with regard to predictive validity:

> A test is considered biased with respect to predictive validity when the inference drawn from the test score is not made with the smallest feasible random error or if there is constant error in an inference or prediction as a function of membership in a particular group.

The above definition of bias is a restatement of previous definitions by a number of researchers and has been widely accepted (though certainly not without criticism, e.g., Bernal, 1975).

The evaluation of bias in prediction under the Cleary et al.

(1975) definition (the regression definition) is quite straightforward. With simple regressions, predictions take the form of $Y_i = \underline{a}X_i + \underline{b}$, where $\underline{a}$ is the regression coefficient and $\underline{b}$ is some constant. When this equation is graphed (forming a regression line), $\underline{a}$ represents the slope of the regression line and $\underline{b}$ the Y-intercept. Since our definition of bias in predictive validity requires errors in prediction to be independent of group membership for the absence of bias, the regression line formed for any pair of variables must be the same for each group for whom predictions are to be made. Whenever the slope or the intercept differs significantly across groups, there is bias in prediction if one attempts to use a regression equation based on the combined groups. When the regression equations for two (or more) groups are equivalent, prediction is the same for all groups. This condition is referred to variously as homogeneity of regression across groups, simultaneous regression, or fairness in prediction. Homogeneity of regression is illustrated in Figure 1, where the single regression equation is appropriate for all groups. That is, errors in prediction from this single

Insert Figure 1 About Here

equation are independent of group membership.

When homogeneity of regression does not occur, there are 3 conditions that can result: a) intercept constants differ, b) regression coefficients (slopes) differ, or c) slopes and

intercepts differ. These conditions are pictured respectively in Figures 2, 3, and 4. Potthoff (1966) has described a useful

Insert Figures 2, 3, and 4 About Here

technique for evaluating homogeneity of regression across groups, allowing one to simultaneously test for equivalence of slopes and intercepts with a single F ratio, that we have used in most of our work.

A considerable body of literature has developed in recent years regarding the differential predictive validity of tests across race for employment selection and college admissions. In a recent review of 866 Black-White test validity comparisons from 39 studies of test bias in personnel selection, Hunter, Schmidt, and Hunter (1979) concluded that there was no evidence to substantiate hypotheses of differential or single-group validity with regard to the prediction of job performance across race for Blacks and Whites. Other racial groupings were not examined by Hunter, Schmidt, and Hunter (1979). A similar conclusion was reached by O'Connor, Wexley, and Alexander (1975). A number of studies have also focused on differential validity of the Scholastic Aptitude Test (SAT) in the prediction of college performance (typically measured by grade point average). In general these studies have found either no difference in the prediction of criterion performance for Blacks and Whites or a bias (underprediction of the criterion) against Whites (see Jensen, 1980, and

Reynolds, 1982, for reviews). When bias against Whites has been found, the differences between actual and predicted criterion scores, while statistically significant, have been quite small. Thus far only one study has been found reporting bias against Blacks in the prediction of college grade point average from SAT scores. The evaluation of bias in the prediction of children's school performance by intelligence tests is more recent however.

Reschly and Sabers (1979) evaluated the validity of WISC-R IQs in the prediction of Metropolitan Achievement Test (MAT) performance (Reading and Math subtests) for Whites, Blacks, Mexican-Americans, and Native American Papagos. The choice of the MAT as a criterion measure in studies of predictive bias is particularly appropriate since item analysis procedures were employed to eliminate racial bias in item content during the test construction phase. Anastasi (1976) points out the MAT as an exemplary model of an achievement test designed to reduce or eliminate cultural bias. The Reschly and Sabers' (1979) comparison of regression systems indicated bias in the prediction of the various achievement scores. Again, however, the bias produced generally significant underprediction of White performance when a common regression equation was applied. Achievement test performance of the Native American Papago group showed the greatest amount of overprediction of all non-White groups. Though some slope bias was evident, Reschly and Sabers typically found intercept bias resulting in parallel regression lines. Using

similar techniques, but including teacher ratings, Reschly and Reschly (1979) also investigated the predictive validity of WISC-R factor scores with samples of White, Black, Mexican-American, and Native American Papago children. A significant relationship occurred between all three WISC-R factors (described earlier) and measures of achievement for the White and non-White groups with exception of the Papagos. Significant correlations occurred between the WISC-R Freedom from Distractibility factor and teacher ratings of attention for all four groups. Reschly and Reschly concluded that "These data also again confirm the relatively strong relationship of WISC-R scores to achievement for most non-Anglo as well as Anglo groups" (p. 359).

Reynolds and Hartlage (1979) investigated the differential validity of Full Scale IQs from the WISC-R and its 1949 predecessor, the WISC, in the prediction of reading and arithmetic achievement for Black and for White children who had been referred by their teachers to psychological services in a rural, southern school district. Comparisons of correlations and a Potthoff analysis to test for identity of regression lines revealed no significant differences in the ability or function of the WISC or WISC-R to predict achievement for these two groups. Reynolds and Nigl (1981) recently replicated this study for groups of black and white, inner city, high poverty district children with the same basic results occurring. These studies were replicated by Reynolds and Gutkin (1980) for the WISC-R with large groups of

White and Mexican-American children from the southwest. Reynolds and Gutkin contrasted regression systems between WISC-R Verbal, Performance, and Full Scale IQs and the "academic basics" of reading, spelling, and arithmetic. Only the regression equation between the WISC-R Performance IQ and arithmetic achievement differed for the two groups. The difference in the two equations was due to an intercept bias that resulted in the overprediction of achievement for the Mexican-American children. Reynolds, Gutkin, Dappen, and Wright (1979) failed to find differential validity in the prediction of achievement for males and females with the WISC-R.

Results with many other individually administered aptitude tests for children consistently have produced similar results. Cross-race comparisons of predictive validity typically reveal no differences across groups whether dealing with school-aged (Bossard, Reynolds, & Gutkin, 1980; Sewell, 1979) or preschool children (Reynolds, 1978, 1980d); when differences do occur, they are in a direction that favors minority groups.

With regard to bias in predictive validity, the empirical evidence suggests conclusions similar to those regarding bias in construct validity. There is no strong evidence to support contentions of differential or single-group validity. Bias occurs infrequently and with no apparently observable pattern, except perhaps with regard to instruments of poor reliability and high specificity of test content. When bias occurs, it is con-

sistently in the direction of favoring low SES, disadvantaged ethnic minority children, or other low scoring groups. Clearly, bias in predictive validity cannot account for the disproportionate number of minority group children diagnosed and placed in EMR or EMH settings.

BIAS IN PERSONALITY SCALES

There is as yet, relatively little study of personality scales that has been cast into the paradigm of test bias. There are conflicting claims on the issue of whether there is cultural bias in personality scales; while Bob Williams claims that entirely different tests are needed to adequately evaluate the personality of blacks, Lewis and her colleagues (as we have previously noted) believe it is discriminatory not to interpret these tests and the behaviors they represent in an equivalent fashion for blacks and whites. It would be our contention that both sides of this issue are without adequate data.

To begin our study of potential bias in personality scales, my colleagues and I have been working with the Revised-Children's Manifest Anxiety Scale (RCMAS) (Reynolds & Richmond, 1978). Thus far, data are available and have been analyzed for three aspects of the problem of bias with this scale: a) item bias (Reynolds, Plake, & Harding, 1981), b) factorial bias (Reynolds & Paget, 1981a), c) bias in the accuracy of measurement (Reynolds & Paget, 1981b). In performing each of these analyses, sex has been included as a group variable along with race. The potential for

cultural bias due to sex is at least as great as that due to race (Reynolds, 1978).

With some minor exceptions, our results thus far are generally commensurate with those for aptitude scales (though we have not yet been able to evaluate differential predictive validity for this scale). Although I have not previously discussed item bias methodology in this presentation, it is necessary and useful to examine for test bias at the individual item level. With an N of nearly 5000 children, we recently completed a study of item bias on the RCMAS for black, white, and Mexican-American children across sex. Using an ANOVA methodology with a Bonferroni-type adjusted follow-up of individual items, a significant race by sex by items interaction was found. Follow-up analyses showed nearly half of the items to be biased in one way or another. Consistent with studies of aptitude tests however, the degree of bias indicated is rather minuscule, the interaction term accounting for less than one percent of the variance of any random observation. The direction of the bias was also counterbalanced across race and sex. Thus any content bias present in the scale appears to be inconsequential at best.

In a just published study (Reynolds & Paget, 1981a), we compared the outcome of separate factor analyses of this scale across race and sex. Five factors were located in the scale and all were found to be congruent across race and sex. Again, the conclusion of factorial similarity was independent of the measure

of similarity employed. Table 3 reports values for male/female

Insert Table 3 About Here

comparisons that indicate a high degree similarity of these 5 factors across sex by each method represented. These values are also quite representative of those produced by the cross race comparisons. Coefficients of congruence (r_c) for the cross race comparison were all above .90, ranging from .91 to .99. A large general anxiety (A_g) factor appeared as well and was highly consistent across race (r_c = .98) and sex (r_c = .99). Thus the factorial validity of the RCMAS, as inferred from this and other (e.g., Reynolds & Richmond, 1979) factor analytic studies, appears to be invariant with regard to race and sex.

In another, just completed, study examining internal consistency estimates across race, sex, and age for the RCMAS, our results are less conclusive (Reynolds & Paget, 1981b). Table 4 presents the alpha reliabilities for the RCMAS at 12 age levels

Insert Table 4 About Here

for white males, white females, black males, and black females. Alpha was compared at each age level for white males vs. black males and for white females vs. black females via a technique described by Feldt (1969) and also discussed by Reynolds (in press). No differences could be detected for males but for females, alpha was significantly lower for blacks than whites at

ages 6, 8, 10, and 11. This was apparently due to some restriction of range at these ages for black females, but nevertheless prompts us to caution against the use of the scale for other than research purposes with black females below the age of 12 years.

Our work with regard to bias in personality assessment must be considered preliminary at present. Much needs to be done. Though the results are thus far promising with regard to the cross-group validity of this scale, and tends to support the position of Lewis and her colleagues, many other scales need to be examined and work with the RCMAS expanded and replicated with other samples. In the meantime however, we must be guided by the existing data.

CONCLUSION

A considerable body of literature currently exists failing to substantiate cultural bias against native born American ethnic minorities with regard to the use of well-constructed, adequately standardized intelligence and aptitude tests. With respect to personality scales, the evidence is promising yet far more preliminary and thus considerably less conclusive. Despite the existing evidence, we do not expect the furor over the cultural test bias hypothesis to be resolved soon. Bias in psychological testing will remain a torrid issue for some time, especially as Larry P. (1979) and PASE (1980) are almost certainly appealed to the U. S. Supreme Court.

Psychologists will need to keep abreast of new findings in

the area however. As new techniques and better methodology are developed and more specific populations examined, the now seen as random, infrequent findings of bias may become better understood and seen to indeed display a correctable pattern.

In the meantime however, psychologists cannot ethically fall prey to the socio-politico legal Zeitgeist of the times and infer bias where none exists. Psychologists cannot justifiably ignore the fact that low IQ, ethnic disadvantaged children are just as likely to fail academically as are their white, middle-class counterparts. Black adolescent delinquents with deviant personality scale scores and exhibiting aggressive behavior need treatment environments as much as their white peers. The potential outcome for score interpretaion, e.g., therapy vs. prison, EMH class vs. regular class, cannot dictate the psychological meaning of test performance. We must practice intelligent testing (Kaufman, 1979). We must remember that it is the purpose of the assessment process to beat the prediction made by the test, to provide insight into hypotheses for environmental interventions that prevent the predicted faliure or subvert the occurance of future maladaptive behavior.

Test developers are also going to have to be more sensitive to the issues of bias, performing appropriate checks for bias prior to test publication. Stereotyping of racial and sexual roles, a fault of many tests that could not be reviewed here, must be halted. Progress is being made in all of these areas. However,

we must hold to the data, even if we do not like it. As my first experimental psychology course professor recited to me as an undergraduate, "the rat is always right." As emotional as the test bias issue has become, we must also be skeptical, even of my talk today, for only in God may we trust, all others must have data.

References

Anastasi, A. Psychological testing. New York: MacMillan, 1976.

Bernal, E. M. A response to "Educational uses of tests with disadvantaged students." American Psychologist, 1975, 30, 93-95.

Blaha, J., Wallbrown, F., & Wherry, R. J. The hierarchical factor structure of the Wechsler Intelligence Scale for Children. Psychological Reports, 1975, 35, 771-778.

Bossard, M. D., Reynolds, C. R., & Gutkin, T. B. A regression analysis of test bias on the Stanford-Binet Intelligence Scale. Journal of Clinical Child Psychology, 1980, 9, 52-54.

Cleary, T. A., Humphreys, L. G., Kendrick, S. A., & Wesman, A. Educational uses of tests with disadvantaged students. American Psychologist, 1975, 30, 15-41.

Dean, R. S. Reliability of the WISC-R with Mexican-American children. Journal of School Psychology, 1977, 15, 267-268.

Dean, R. S. WISC-R factor structure for Anglo and Hispanic children. Paper presented to the annual meeting of the American Psychological Association, New York, September, 1979.

DeFries, J. C., Vandenberg, S. G., McClearn, G. E., Kuse, A. R., Wilson, J. R., Ashton, G. C., & Johnson, R. C. Near identity of cognitive structure in two ethnic groups. Science, 1974, 183, 338-339.

Feldt, L. S. A test of the hypothesis that Cronbach's alpha or Kuder-Richardson coefficient twenty is the same for two tests. Psychometrika, 1969, 34, 363-373.

Gutkin, T. B., & Reynolds, C. R. Factorial similarity of the WISC-R for Anglos and Chicanos referred for psychological services. Journal of School Psychology, 1980, 18, 34-39.

Gutkin, T. B., & Reynolds, C. R. Factorial similarity of the WISC-R for white and black children from the standardization sample. Journal of Educational Psychology, 1981, 73, 227-231.

Hilliard, A. G. Standardization and cultural bias as impediments to the scientific study and validation of "intelligence." Journal of Research and Development in Education, 1979, 12, 47-58.

Hirsch, J. To "unfrock the charlatans." SAGE Race Relations Abstracts, 1981, 6, 1-67.

Horn, J., & Goldsmith, H. Reader be cautious: A review of Bias in Mental Testing. American Journal of Education, 1981, 89, 305-329.

Hunter, J. E., Schmidt, F. L., & Hunter, R. Differential validity of employment tests by race: A comprehensive review and analysis. Psychological Bulletin, 1979, 86, 721-735.

Jensen, A. R. How biased are culture-loaded tests? Genetic Psychology Monographs, 1974, 90, 185-224.

Jensen, A. R. An examination of bias in the Wonderlic Personnel Test. Intelligence, 1977, 1, 51-64.

Jensen, A. R. Bias in mental testing. New York: The Free Press, 1980. a.

Jensen, A. R. Correcting the bias against mental tests: A preponderance of peer agreement. Behavioral and Brain Sciences, 1980, 3, 359-371. b

Kaufman, A. S. Intelligent testing with the WISC-R. New York: Wiley-Interscience, 1979.

Kaufman, A. S., & DiCuio, R. Separate factor analyses of the McCarthy Scales for groups of black and white children. Journal of School Psychology, 1975, 13, 10-18.

Kaufman, A. S., & Hollenbeck, G. P. Comparative structure of the WPPSI for blacks and whites. Journal of Clinical Psychology, 1974, 30, 316-319.

Larry P. et al. v. Riles et al. No. C 71 2270. U.S.D.C. for the Northern District of California, San Francisco, October, 1979, slip opinion.

Lewis, D. O., Balla, D. A., & Shanok, S. S. Some evidence of race bias in the diagnosis and treatment of the juvenile offender. American Journal of Orthopsychiatry, 1979, 49, 53-61.

Lewis, D. O., Shanok, S. S., Cohen, R. J., Kligfield, M., & Frisone, G. Race bias in the diagnosis and disposition of violent adolescents. American Journal of Psychiatry, 1980, 137, 1211-1216.

Lindsey, J. The factorial organization of intelligence as related to the variables of age, sex, and subculture. Unpublished doctoral dissertation, University of Georgia, Athens, 1967.

Mercer, J. R. What is a racially and culturally non-discriminatory test? A sociological and pluralistic perspective. In C. R. Reynolds & R. T. Brown (Eds.) Perspectives on Bias in Mental Testing, New York: Plenum, in press.

Merz, W. R. A factor analysis of the Goodenough-Harris Drawing Test across four ethnic groups. (Doctoral dissertation, University of New Mexico, 1970). Dissertation Abstracts International, 1970 (University Microfilms #70-19, 714).

Miele, F. Cultural bias in the WISC. Intelligence, 1979, 3, 149-164.

Mulaik, S. A. The foundations of factor analysis. New York: McGraw-Hill, 1972.

Oakland, T., & Feigenbaum, D. Multiple sources of test bias on the WISC-R and the Bender-Gestalt test. Journal of Consulting and Clinical Psychology, 1979, 47, 968-974.

O'Conner, E. J., Wexley, K. N., & Alexander, R. A. Single group validity: Fact or fallacy. Journal of Applied Psychology, 1975, 60, 352-355.

PASE: Parents in action on special education et al. v. Hannon et al. No. 74 C 3586. U.S.D.C. for the Northern District of Illinois, Eastern Division, Chicago, July, 1980, slip opinion.

Potthoff, R. F. Statistical aspects of the problem of bias in psychological tests. Institute of Statistics Mimeo Series No. 479, Chapel Hill, NC: University of North Carolina Department of Statistics, 1966.

Reschly, D. J. WISC-R factor structures among Anglos, Blacks, Chicanos, and Native American Papagos. Journal of Consulting and Clinical Psychology, 1978, 46, 417-422.

Reschly, D. J., & Reschly, J. E. Validity of WISC-R factor scores in predicting achievement and attention for four sociocultural groups. Journal of School Psychology, 1979, 17, 355-361.

Reschly, D. J., & Sabers, D. Analysis of test bias in four groups with the regression definition. Journal of Educational Measurement, 1979, 16, 1-9.

Reynolds, C. R. Differential validity of several preschool assessment instruments for blacks, whites, males, and females. Unpublished doctoral dissertation, University of Georgia, Athens, 1978.

Reynolds, C. R. The invariance of the factorial validity of the Metropolitan Readiness Tests for blacks, whites, males, and females. Educational and Psychological Measurement, 1979, 39, 1047-1052.

Reynolds, C. R. In support of "Bias in Mental Testing" and scientific inquiry. Behavioral and Brain Sciences, 1980, 3, 352. a.

Reynolds, C. R. Differential construct validity of a preschool battery for blacks, whites, males, and females. Journal of School Psychology, 1980, 18, 112-125. b.

Reynolds, C. R. Differential construct validity of intelligence as popularly measured: Correlations of age with raw scores on the WISC-R for blacks, whites, males, and females. Intelligence, 1980, 4, 371-379. c.

Reynolds, C. R. An examination for test bias in a preschool battery across race and sex. Journal of Educational Measurement, 1980, 17, 137-146. d.

Reynolds, C. R. The problem of bias in psychological assessment. In C. R. Reynolds & T. B. Gutkin (Eds.), The Handbook of School Psychology, New York: Wiley, 1982.

Reynolds, C. R. Methods for detecting construct and predictive bias. In R. A. Berk (Ed.), Handbook of Methods for Detecting Test Bias, Baltimore: Johns Hopkins University Press, in press.

Reynolds, C. R., & Brown, R. T. Bias in mental testing: An introduction to the issues. In C. R. Reynolds & R. T. Brown (Eds.), Perspectives on Bias in Mental Testing, New York: Plenum, in press. a.

Reynolds, C. R., & Brown, R. T. (Eds.) Perspectives on bias in mental testing. New York: Plenum, in press. b.

Reynolds, C. R., Gutkin, T. B., Dappen, L., & Wright, D. Differential validity of the WISC-R for boys and girls referred for psychological services. Perceptual and Motor Skills, 1979, 48, 868-870.

Reynolds, C. R., & Gutkin, T. B. A regression analysis of test bias on the WISC-R for Anglos and Chicanos referred to psychological services. Journal of Abnormal Child Psychology, 1980, 8, 237-243.

Reynolds, C. R., & Harding, R. E. Comparative outcome in two studies of factorial similarity under six methods of factor comparison. Manuscript under review, 1981.

Reynolds, C. R., & Hartlage, L. C. Comparison of WISC and WISC-R regression lines for academic prediction with black and with white referred children. Journal of Consulting and Clinical Psychology, 1979, 47, 589-591.

Reynolds, C. R., & Nigl, A. J. A regression analysis of differential validity of intelligence tests for black and for white inner city children. Journal of Clinical Child Psychology, 1981, in press.

Reynolds, C. R., & Paget, K. D. Factor analysis of the revised children's manifest anxiety scale for blacks, whites, males, and females with a national normative sample. Journal of Consulting and Clinical Psychology, 1981, 49, 352-359. a.

Reynolds, C. R., & Paget, K. D. National normative and reliability data for the Revised-Children's Manifest Anxiety Scale for blacks, whites, males, and females at 12 age levels. Manuscript in preparation, 1981, b.

Reynolds, C. R., & Piersel, W. C. Item homogeneity estimates on the Boehm Test of Basic Concepts (Forms A & B) for whites and Mexican-Americans. Manuscript in review. 1981.

Reynolds, C. R., Plake, B. S., & Harding, R. E. Group by item interactions in the assessment of children's anxiety levels: Effects of race, sex, and age on item responses. Manuscript under review, 1981.

Reynolds, C. R., & Richmond, B. O. What I Think and Feel: A revised measure of children's manifest anxiety. Journal of Abnormal Child Psychology, 1978, 6, 271-280.

Reynolds, C. R., & Richmond, B. O. Factor structure and construct validity of "What I Think and Feel:" The Revised-Children's Manifest Anxiety Scale. Journal of Personality Assessment, 1979, 43, 281-283.

Sandoval, J. The WISC-R and internal evidence of test bias with minority children. Journal of Consulting and Clinical Psychology, 1979, 47, 919-927.

Sandoval, J., Zimmerman, I. L., & Woo-Sam, J. Cultural differences on WISC-R verbal items. Paper presented to the annual meeting of the American Psychological Association, Montreal, September, 1980.

Schoenfeld, W. N. Notes on a bit of psychological nonsense: "Race differences in intelligence." Psychological Record, 1974, 24, 17-32.

Semler, I., & Iscoe, I. Structure of intelligence in Negro and White children. Journal of Educational Psychology, 1966, 57, 326-336.

Sewell, T. E. Intelligence and learning tasks as predictors of scholastic achievement in black and white first-grade children. Journal of School Psychology, 1979, 17, 325-332.

Silverstein, A. B. Factor structure of the Wechsler Intelligence Scale for Children for three ethnic groups. Journal of Educational Psychology, 1973, 65, 408-410.

Vance, H. B., & Wallbrown, F. H. The structure of intelligence for black children: A hierarchical approach. The Psychological Record, 1978, 28, 31-39.

Wallbrown, F. H., Blaha, J., Wallbrown, J., & Engin, A. The hierarchical factor structure of the Wechsler Adult Intelligence Scale. British Journal of Educational Psychology, 1974, 44, 47-55.

Wechsler, D. Wechsler Intelligence Scale for Children-Revised. New York: The Psychological Corporation, 1974.

Williams, R. L. From dehumanization to black intellectual genocide: A rejoinder. In G. J. Williams & Sol Gordon (Eds.), Clinical Child Psychology: Current Practices and Future Perspectives, New York: Behavioral Publications, 1974.

Table 1

Indexes of Factorial Similarity for Three-Factor Solut ons of the Wechsler Intelligence Scale for Children-Revised for Blacks and Whites

	Index of Similarity							
	Coefficient of Congruence, Correlation Matrix	Coefficient of Congruence, Covariance Matrix	Pearson r, No Transformation	Pearson r, Fisher Transformation	Salient Variable Similarity Index[a] .10	.20	Factor Score Correlation[b] Blacks	Whites
Factor								
1. Verbal Comprehension	.99	.99	.98	.98	.96	88	.99	.99
2. Perceptual Organization	.99	.99	.95	.95	1.00	.91	.99	.99
3. Freedom from Distractibility	.99	.98	.96	.94	.91	.88	.98	.99

[a]Reported using two separate cutoff values to indicate salience, .10 as recommended by Cattell (1978) and .20 as suggested by Cattell when a conservative stance is taken.

[b]Correlations for Blacks are reported between scores for each individual based on factor scores derived from formulas based on a Black only analysis and scores from a total sample analysis. White correlations are from a white only analysis compared to a total sample analysis.

Table 2

Coefficients of Congruence Between WISC-R Factors Derived from the Scores of Whites in the Standardization Sample and Factors From Studies Reporting Two- and Three-Factor WISC-R Solutions for Blacks and Mexican-Americans.

		Coefficient of Congruence	
"g" factor:		Median	Range
Blacks		.99	.98 - .99
Mexican-Americans		.99	.97 - .99
Two-Factor Solutions:			
Factor 1	Blacks	.96	.93 - .99
	Mexican-Americans	.97	.96 - .99
Factor 2	Blacks	.96	.94 - .99
	Mexican-Americans	.96	.95 - .99
Three-Factor Solutions:			
Factor 1	Blacks	.98	.95 - .99
	Mexican-Americans	.95	.94 - .98
Factor 2	Blacks	.93	.90 - .99
	Mexican-Americans	.92	.81 - .96
Factor 3	Blacks	.94	.71 - .99
	Mexican-Americans	.91	.61 - .96

Table 3

Indexes of Factorial Similarity for Five-Factor Solutions of the Revised-Children's Manifest Anxiety Scale for Males and Females

	Index of Similarity							
	Coefficient of Congruence, Correlation Matrix	Coefficient of Congruence, Covariance Matrix	Pearson r, No Transformation	Pearson r, Fisher Transformation	Salient[a] Variable Similarity Index .10	.20	Factor[b] Score Correlation Males	Females
Factor								
1. Physiological	.99	.98	.99	.99	.98	.83	.99	.99
2. Worry/Oversensitivity	.99	.95	.95	.95	.96	.94	.99	.99
3. Concentration	.96	.94	.96	.97	.92	.94	.99	.99
4. Lie I	.99	.95	.90	.90	.84	1.00	.98	.97
5. Lie II	.99	.96	.97	.98	.75	1.00	.99	.99

[a]Reported using two separate cutoff values to indicate salience, .10 as recommended by Cattell (1978) and .20 as suggested by Cattell when a conservative stance is taken.

[b]Correlations for males are reported between scores for each individual based on factor scores derived from formulas based on a male only analysis and scores from a total sample analysis. Female correlations are from a female only analysis compared to a total sample analysis.

Table 4

Internal Consistency Estimates for the RCMAS Total Anxiety Scale Score Reported for White Males, White Females, Black Males, and Black Females at 12 Age Levels.

Coefficient Alpha Reliability Estimates

Age Level	White Males[a]	White Females	Black Males	Black Females
6	.78 (70)	.84 (90)	.83 (15)	.42* (11)
7	.78 (200)	.79 (183)	.84 (32)	.76 (45)
8	.80 (261)	.78 (254)	.80 (47)	.66* (51)
9	.83 (291)	.81 (262)	.82 (42)	.76 (45)
10	.80 (246)	.85 (234)	.77 (30)	.70* (35)
11	.83 (250)	.85 (250)	.85 (31)	.75* (36)
12	.82 (176)	.86 (175)	.87 (34)	.79 (26)
13	.84 (98)		.75 (9)	.86 (10)
14	.83 (80)	.82 (75)	.81 (21)	.62 (6)
15	.83 (168)	.81 (170)	.87 (10)	.80 (5)
16	.82 (122)	.78 (140)	.84 (6)	.82 (8)
17-19	.78 (243)	.79 (261)	.87 (11)	.87 (20)

[a] N in parentheses

*Significantly lower than corresponding value for white females.
No differences occurred between White Males and Black Males.

Figure Captions

Figure 1. Equal slopes and intercepts result in homogeneity of regression that causes the regression line for group a, group b, and the common regression line derived by combining the two groups to be identical.

Figure 2. Equal slopes with differing intercepts result in parallel regression lines and a constant bias in prediction.

Figure 3. Equal intercepts and differing slopes result in nonparallel regression lines with the degree of bias dependent on the distance of the score (X_i) from the origin.

Figure 4. Differing slopes and intercepts result in the complex condition where the degree and direction of bias is a function of the distance of the score (X_i) from the origin.

Fig. 2

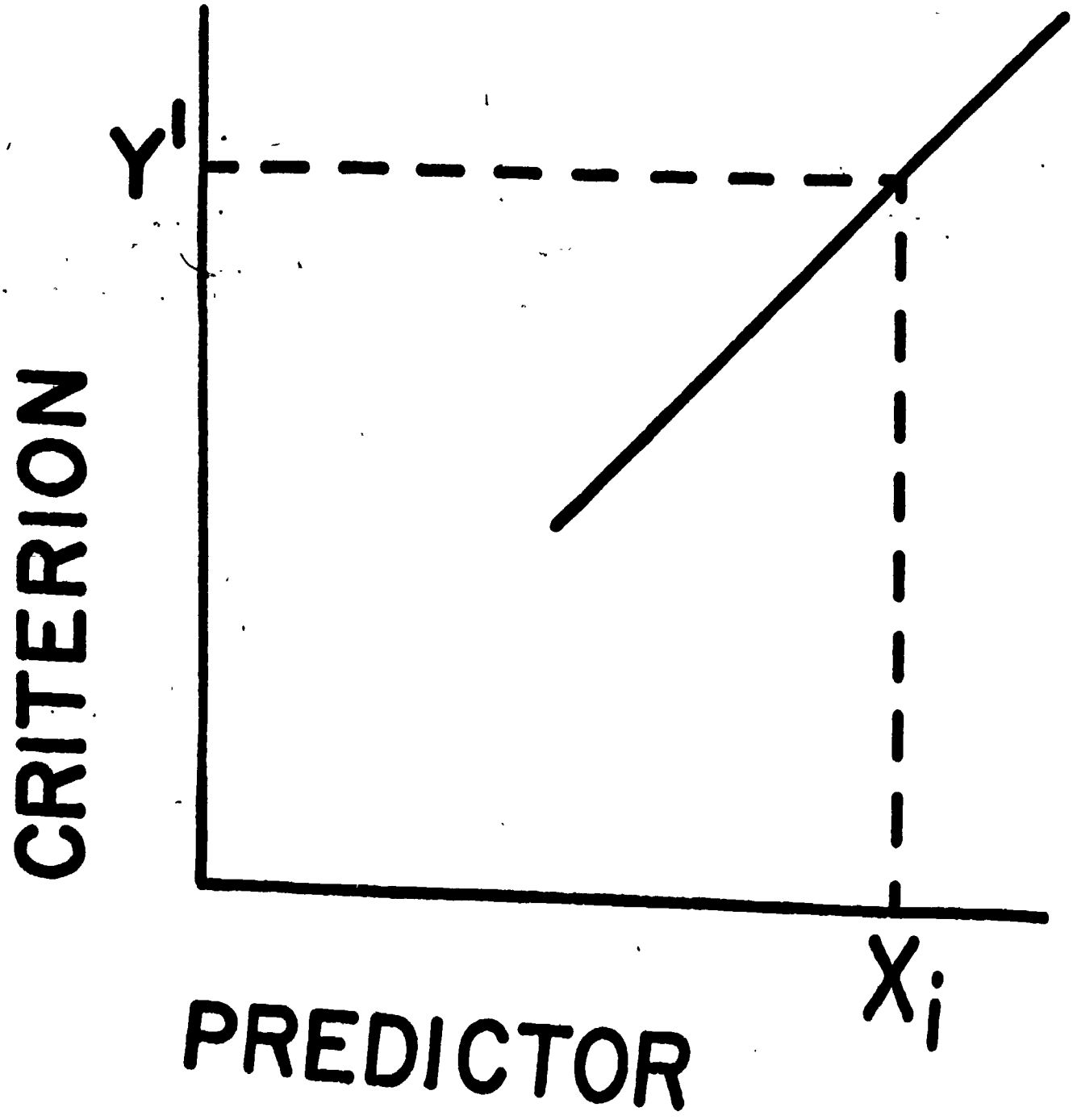


Fig. 2

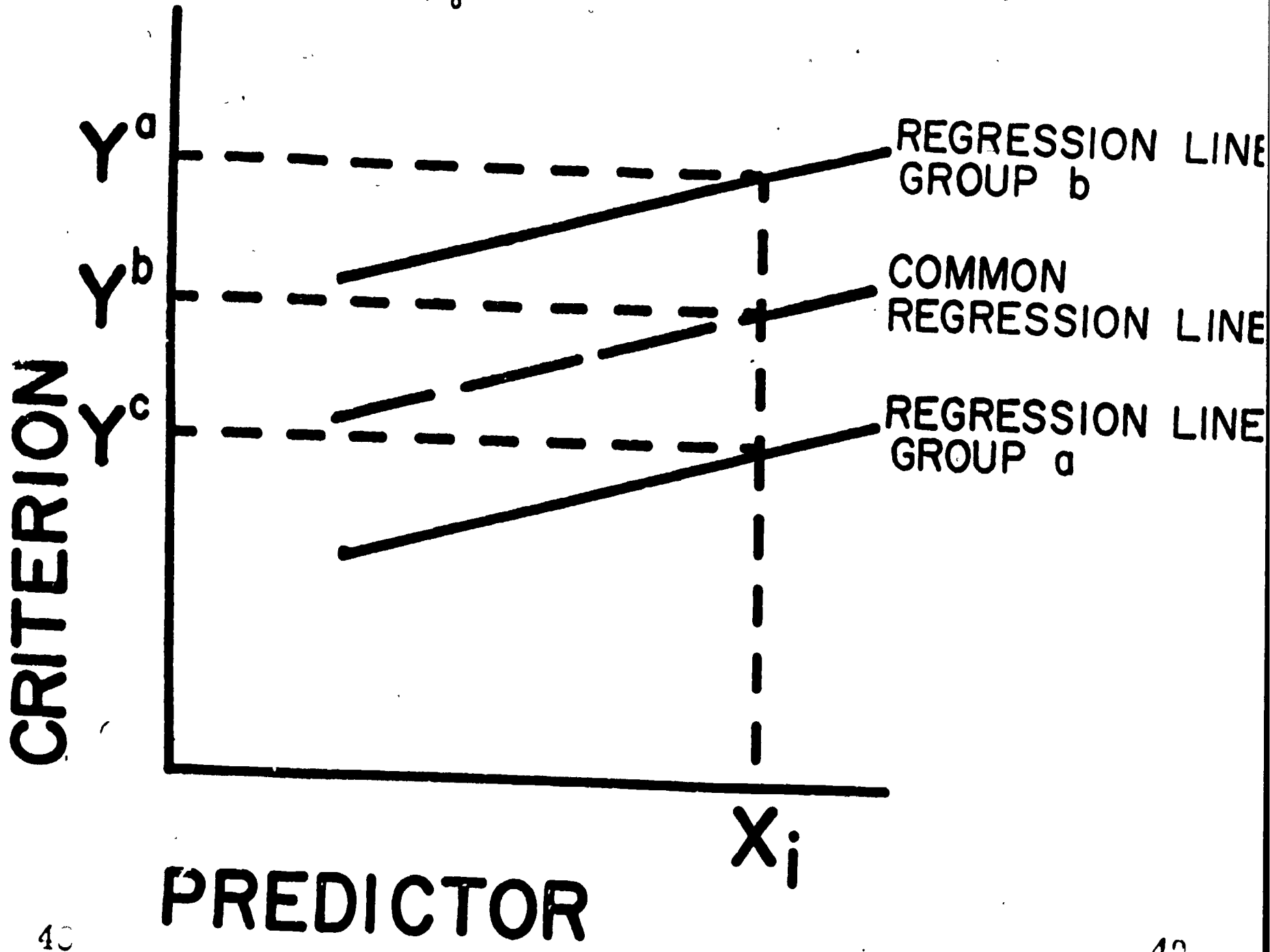
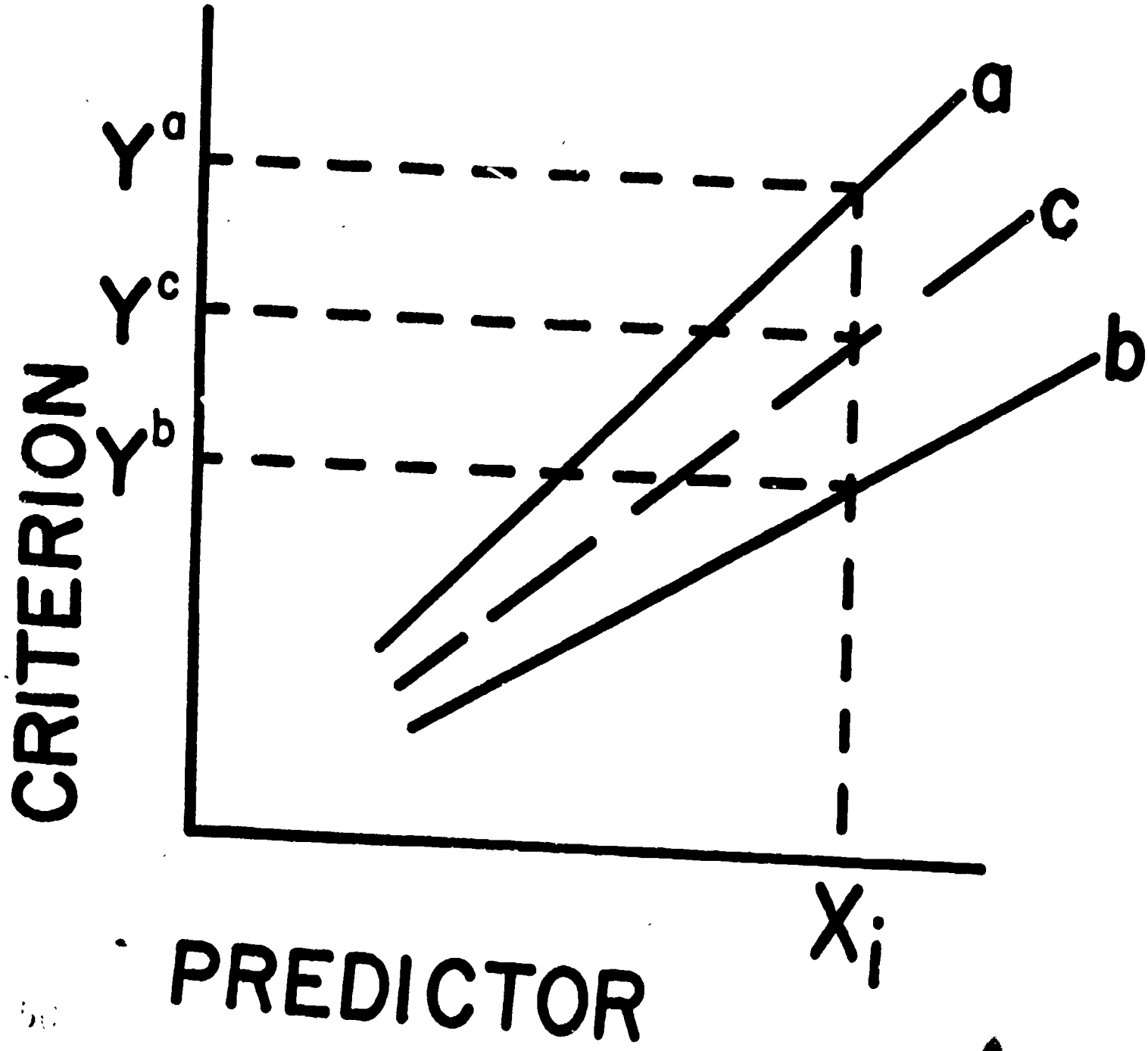

Y^a
Y^b
Y^c
REGRESSION LINE
GROUP b
COMMON
REGRESSION LINE
REGRESSION LINE
GROUP a
CRITERION
X_i
PREDICTOR

Fig. 3



F.8.4