Regressions lines for the prediction of Wide Range Achievement Test (WRAT) standard scores by Stanford-Binet Intelligence Scale scores were compared across race for matched groups of 60 black and 60 white children selected from among a large number of children who had been referred for psychological services by their classroom teachers. The white children were matched to the black children on the basis of sex, age, and IQ. The definition proposed for bias was significant differences in regression lines using Potthoff's technique that tests for slopes and intercepts simultaneously. According to this significance test, regression lines for blacks and for whites did not differ significantly for the prediction of WRAT scores. (Author/CTM)
A Regression Analysis of Test Bias on the Stanford-Binet Intelligence Scale

Cecil R. Reynolds
Acting Director
Buros Institute of Mental Measurements
135 Bancroft Hall
University of Nebraska-Lincoln
Lincoln, Nebraska 68588

Michael D. Bossard
Doctoral Candidate
School Psychology Training Program
130 Bancroft Hall
University of Nebraska-Lincoln
Lincoln, Nebraska 68588

Terry B. Gutkin
Director
School Psychology Training Program
130 Bancroft Hall
University of Nebraska-Lincoln
Lincoln, Nebraska 68588

Abstract

Regression lines for the prediction of WRAT standard scores by Stanford-Binet IQs were compared across race, by the Potthoff procedure, for equated groups (sex, age, and IQ) of 60 Black and 60 White children referred for psychological services by their classroom teachers. Regression lines for Blacks and Whites did not differ significantly for the prediction of WRAT scores by the Stanford-Binet. Implications of these findings are discussed.
The use of psychological tests normed primarily with white children for psychological diagnosis and educational decision-making concerning minority children, has become a very hotly contested issue in recent years. Although much discussion of the issue has appeared in both the scientific and public literature, few data of relevance to the issue (with school age children) have been presented. The use of such tests are of special concern to psychologists involved in assessment, particularly in view of the Larry P. case (Note 1) and P. L. 94-142 (Note 2; Education for All Handicapped Children Act of 1975). Harrington (1976, 1975) has gone so far as to state that it is not possible for tests developed and normed on a white majority to be other than biased against minorities and to show less predictive validity when used with minorities.

In response to pressure from the Black Psychological Association (which was actually requesting a moratorium on the use of psychological and educational tests with disadvantaged students), the APA Board of Directors requested, in 1968, the Board of Scientific Affairs to appoint a group to study the use of such tests with disadvantaged students. In reporting on this issue, the committee (Cleary, Humphreys, Kendrick, & Wesman, 1975) offered a definition of test bias. While including content and construct validity as important variables in the issue of test bias, the focus was clearly
A test is considered fair for a particular use if the inference drawn from the test score is made with the smallest feasible random error and if there is no constant error in the inference as a function of membership in a particular group. (Cleary et al., 1975, p.25)

The definition of bias offered by the APA committee is a restatement of previous definitions by Cardall and Coffman (1964), Cleary (1968), and Potthoff (1966), and has been widely accepted (though certainly not without criticism, e.g., Bernal, 1975; Linn & Werts, 1971; Thorndike, 1971). Oakland and Matuszek (1977) examined class placement procedures under several proposed models of bias and demonstrated that the Cleary model results in the smallest number of children being misplaced, although under certain legislative conditions, Oakland and Matuszek favored the Thorndike (1971) "quota" selection model. After reviewing a number of models, Peterson and Novick (1976) designated the regression model as the most logically tenable and the most widely used placement model. A statistical technique provided by Potthoff (1966) has also received widespread acceptance in the examination of regression lines to test bias under the Cleary et al. definition (Schmidt & Hunter, 1974).

While considerable data are available on the validity of the Scholastic Achievement Test (e.g., Goldman & Hewitt, 1976; Kallingal, 1971; Pfeifer & Sedlacek, 1971) and various employment tests (e.g. Boehm, 1972; Hunter, Schmidt, & Hunter, 1979) for
blacks and whites, only recently have studies appeared dealing with differential validity of IQ tests. Mitchell (1967) studied the validity of two broad based readiness tests to predict first grade achievement for blacks and whites finding similar validity coefficients for the two races. Mitchell's study was limited to comparing the magnitude of independent-dependent variable correlation and did not look for identity of regression lines. Hartlage, Lucas, and Godwin (1976) compared the predictive validity of the WISC and Raven with a group of low SES, disadvantaged children. When comparing what they considered to be the relatively culture-fair test, the Raven Matrices, with the "culture-loaded" 1949 WISC, Hartlage et al. (1976) found the WISC to have consistently larger correlations with measures of reading, spelling, and arithmetic than the Raven. These authors only compared the strength of the relationship in each case and did not look for identity of regression lines (equivalent beta coefficients and intercept constants).

More recently, Reynolds and Hartlage (1979) compared regression lines for the prediction of achievement by the WISC and the WISC-R across race for blacks and whites. Their results indicated that regression lines for blacks and whites did not differ significantly. Reynolds and Gutkin (1980) replicated the Reynolds and Hartlage (1979) study for the WISC-R, comparing regression lines between whites and Mexican-Americans. Again, no significant differences
were found. In a study with much larger samples, Reschly and Sabers (1979) investigated the WISC-R's ability to predict Metropolitan Achievement Test scores across four ethnic groups (blacks, whites, chicanos, and native American Papagos). Reschly and Sabers (1979) adopted the Cleary, regression definition and a procedure by Gulliksen and Wilks (1950) that separately tests slopes and intercepts (whereas the Potthoff, 1966, technique simultaneously tests slopes and intercepts). They found that the WISC-R was for the most part equally valid for the different groups. When differences occurred, they were due to variations in intercepts resulting in the over-prediction of performance for non-white groups.

The purpose of the present study is to provide data that will aid in the empirical evaluation of test bias (under the Cleary et al., 1975, definition) for the Stanford-Binet Intelligence Scale, Form L-M, 1972 Norms Edition (Terman & Merrill, 1973). It was hypothesized that, as with previous research on the WISC and WISC-R, no significant differences would occur between regression lines across groups. Previous research on bias has ignored the Binet. The Binet should be of particular interest in test bias research since it has historically been the IQ test against which new tests have been validated.
METHOD

Subjects

The sample consisted of equated groups of 60 white and 60 black urban children referred by teachers for psychological evaluation due to a variety of learning and/or behavior problems. A referral population was chosen because they are the predominant group of interest in the prediction of achievement from the IQ. The children were chosen as follows from more than 1,000 district-wide referrals. A computer listing of all children with complete data was obtained. Every third black male was chosen until 30 children were obtained. The procedure was repeated for black females. Since random assignment to race or sex is not possible, whites were chosen to match the black children on the variables of age (within 6 months), sex, and IQ (within 10 points). To match the groups, a black child was chosen and records of the white group examined. The first matching white child to be encountered was selected. The resulting sample characteristics are described in greater detail in Table 1. The relatively low IQ of the groups is typical of referral populations (Gutkin & Reynolds, in press; Reynolds, Gutkin, Dappen, & Wright, 1979; Reynolds & Hartlage, 1979).

---

Insert Table 1 about here
Procedure

The Stanford-Binet Intelligence Scale (Terman & Merrill, 1973) and the most recent revision of the Wide Range Achievement Test (Jastak & Jastak, 1978) were administered by certified school psychologists and psychological assistants. Testing on both scales was accomplished during a single session.

Regression lines for each pair of scores (Binet IQ predicting each WRAT subtest) were examined across race through the Potthoff (1966) technique. This procedure yields a single $F$-ratio that simultaneously tests regression coefficients (slopes) and intercept values. If a significant $F$ results, slopes and intercepts may then be assessed separately to determine whether the resulting bias in prediction is constant (intercepts differ) or changes with the distance of scores from the mean (slopes differ). Slopes and intercepts must both be equivalent prior to concluding homogeneity of regression across groups. Only when slope and intercepts are the same can a common regression equation (derived by combining the groups in question) be applied. If homogeneity of regression across groups does not occur, then in order to have fair use of test scores, separate equations for each group must be employed.

RESULTS

Regression lines for blacks and whites did not differ at the .05 level of significance for the prediction of WRAT Reading, $F(2,116) = 1.24$, $p > .05$, Spelling, $F(2,116) = 0.18$, $p > .05$, or
Arithmetic, F (2,116) = 2.24 p > .05, standard scores by the Stanford-Binet IQ. Thus, present results provide support for the use of a common regression equation (Bossard & Galusha, in press) to predict WRAT achievement scores for referred black and white children with the Stanford-Binet. Correlations between the Stanford-Binet IQ and achievement for both groups were quite substantial, never accounting for less than 49% of the variance in achievement scores. For black children the correlations were: .74 with Reading, .78 with Spelling, and .70 with Arithmetic. For whites the correlations were: .81 with Reading, .81 with Spelling, and .82 with Arithmetic. As expected from the results of the Pott-hoff analysis, the pairs of correlations are quite similar across these two racial groupings.

DISCUSSION

The study’s results are consistent with previous investigations of test bias using the regression definition. That is, standardized intelligence tests have been shown to predict school achievement about equally well for blacks and whites. Prior to concluding that the Stanford-Binet Intelligence Scale is free of bias in terms of predictive accuracy (the regression definition), more research is needed utilizing a wide variety of criterion measures including other individual achievement tests, group achievement tests, and teacher constructed scales. Studies of this kind will help to evaluate the relative influence of bias within different criterion
measures. Since using a referral population may minimize differences between groups, replication with normal children will also need to be undertaken.

Test developers need to become more aware of the issue of bias, to the point of demonstrating validity across groups prior to publication of the instrument. While this has occurred somewhat in the area of achievement testing (Anastasi, 1976), investigations of differential validity by test publishers are conspicuously lacking. Studies similar to the present investigation are needed with other existing measurement instruments to determine whether alterations in interpretation of the scales are needed when applied to groups other than the majority population.

At present however, a considerable body of data is accumulating indicating consistency of content (Jensen & Figueroa, 1975), construct (Gutkin & Reynolds, in press; Jensen, 1976; Reschly, 1978; Reynolds, in press a,b), and predictive (Reschly & Sabers, 1979; Reynolds & Gutkin, 1980; Reynolds & Hartlage, 1979) validity of the IQ test across racial groupings.
Reference Notes


References


Goldman, R. & Hewitt, B. Predicting the success of black, chicano, oriental, and white college students. Journal of Educational
Gulliksen, J., & Wilks, S. Regression tests for several samples.
Psychometrika, 1950, 15, 91-114.
Gutkin, T. B., & Reynolds, C. R. Factorial similarity of the WISC-R for Anglos and Chicanos referred for psychological services.
Journal of School Psychology, in press.


### Table 1

**Sample Characteristics by Race and Sex**

<table>
<thead>
<tr>
<th>Sex</th>
<th>N</th>
<th>Age in Years</th>
<th>Stanford-Binet IQ</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>X</td>
<td>SD</td>
</tr>
<tr>
<td>Blacks</td>
<td>M</td>
<td>30</td>
<td>8.38</td>
</tr>
<tr>
<td></td>
<td>F</td>
<td>30</td>
<td>8.53</td>
</tr>
<tr>
<td>Whites</td>
<td>M</td>
<td>30</td>
<td>8.30</td>
</tr>
<tr>
<td></td>
<td>F</td>
<td>30</td>
<td>8.42</td>
</tr>
</tbody>
</table>

**Wide Range Achievement Test**

<table>
<thead>
<tr>
<th>Sex</th>
<th>N</th>
<th>Reading</th>
<th>Spelling</th>
<th>Arithmetic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>X</td>
<td>SD</td>
<td>X</td>
</tr>
<tr>
<td>Blacks</td>
<td>M</td>
<td>30</td>
<td>83.43</td>
<td>16.11</td>
</tr>
<tr>
<td></td>
<td>F</td>
<td>30</td>
<td>83.30</td>
<td>16.47</td>
</tr>
<tr>
<td>Whites</td>
<td>M</td>
<td>30</td>
<td>82.97</td>
<td>16.63</td>
</tr>
<tr>
<td></td>
<td>F</td>
<td>30</td>
<td>84.90</td>
<td>23.48</td>
</tr>
</tbody>
</table>