DOCUMENT RESUME

ED 126 111 TM 005 366

AUTHOR Beschly, Daniel J.; And Others
TITLE Analysis of Different Concepts of Cultural Fairness Using WISC-R and SAT Scores from Four Ethnic Groups.
SPONS AGENCY Iowa State Univ. of Science and Technology, Ames. Sciences and Humanities Research Inst.
PUB DATE [Apr 76]
NOTE 17p.; Paper presented at the Annual Meeting of the American Educational Research Association (60th, San Francisco, California, April 19-23, 1976)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS *Achievement Tests; American Indians; Caucasian Students; Comparative Analysis; Cross Cultural Studies; Elementary Secondary Education; *Ethnic Groups; *Intelligence Tests; Mexican Americans; Negro Students; Norms; Predictive Ability (Testing); *Standardized Tests; Statistical Analysis; *Test Bias; Test Interpretation
IDENTIFIERS *Metropolitan Achievement Test; *Wechsler Intelligence Scale for Children Revised

ABSTRACT

Samples of Anglo, Black, Mexican-American, and Papago Indian students (N=260 per group) were administered the Metropolitan Achievement Test and the Wechsler Intelligence Scale for Children-Revised (WISC-R). Various concepts of cultural fairness were identified from the literature and then used to design analyses of these data. Regression equations for each group were formulated and compared. Suggestions for interpretation of standardized test scores among culturally different students are provided on the basis of these data on cultural fairness. The formulation, application, and implications of pluralistic norms are discussed. (Author/RC)

ANALYSIS OF DIFFERENT CONCEPTS OF CULTURAL FAIRNESS USING WISC-R AND MAT SCORES FROM FOUR ETHNIC GROUPS[1]

Daniel J. Reschly
Iowa State University

Darrell L. Sabers
University of Arizona

Keith E. Meredith
University of Arizona

The recent HEW Conference in Atlanta on nondiscriminatory testing and the still pending court suit in San Francisco (Larry P. vs Wilson Riles) regarding the use of individual intelligence tests are but two of the many recent expressions of concern about the use of standardized tests with minority students. In the Larry P. case the preliminary hearing resulted in an injunction suspending the use of individual intelligence tests with Black students in California. In this court case as well as in other contexts, decisions about future test use were based on the notion of test bias, a concept which is subject to a variety of interpretations.

The current literature on test bias includes highly diverse opinions on this crucial issue. Three general conceptualizations of test bias seem especially prominent. One point of view indicts standardized tests as biased whenever mean differences in performance are found among different groups. Specifically, tests are defined as biased if different racial and/or ethnic groups obtain scores which on the average are below population means. Advocates of this point of view (e.g., Jackson, 1975; Williams, 1974) emphasize a

---

[1]Paper delivered at the American Educational Research Association Annual Meeting, April, 1976, San Francisco.

Presentation of this paper was supported in part by the Sciences and Humanities Research Institute of Iowa State University.

discrepancy in test content and cultural background and/or the influence of testing atmosphere (e.g., examiner, setting) as the major elements of bias. Almost all current standardized tests are characterized as not only biased, but also unfair and discriminatory. Revisions of present tests in the direction of greater cultural specificity and/or complete abolition of current testing practices are frequently suggested as remedies.

A second position on the test bias concept stresses the use of standardized tests in predicting academic achievement and/or success in employment settings. From this point of view a test may be defined as biased or unbiased depending upon the effectiveness and accuracy of prediction for all groups of concern. Even though mean differences in performance on the test may exist, test use is regarded as unbiased if the test can be shown to be an accurate and "fair" predictor ("fair" in this sense means that the same criterion score is predicted for individuals obtaining identical test scores regardless of group membership). The following definition of test bias formulated by Cleary has been frequently cited by investigators pursuing this line of reasoning.

> A test is biased for members of a subgroup of the population if, in the prediction of a criterion for which the test is designed, consistent nonzero errors of prediction are made for members of the subgroup. In other words, the test is biased if the criterion score predicted from the common regression line is consistently too high or too low for members of the subgroup. With this definition of bias, there may be a connotation of 'unfair,' particularly if the use of the test produces a prediction that is too low. (Cleary, 1968, p. 115)

Considerable research on test use was stimulated and guided by the Cleary definition (e.g., Boehm, 1972; Cleary, 1968; Kallingal, 1971; Pfeifer & Sedlacek, 1971; Schmidt, Berner, & Hunter, 1973). The findings which emerged from these studies indicated that tests were approximately equally valid for Black and White groups at least in the college and employment settings studied.

Moreover, errors of prediction tended to be in the direction of overprediction for Blacks and underprediction for Whites when the same regression equation was applied to both groups.

Alternatives to the Cleary definition of test bias were suggested by Thorndike (1971) and Cole (1973). These definitions continued the stress on prediction accuracy and implied examination of regression planes as an important component of attempts to assess degree of bias in test use.

A third conception of test bias in the literature recognizes the social policy implications of test use. The fact that current tests predict accurately for diverse groups does little or nothing in terms of reducing the historical inequities among the groups. Darlington (1971) suggested recognition of the social policy issues inherent in any discussion of test bias. The culturally sensitive position advocated by Darlington and recently amplified and stated formally by Peterson and Novick (1976) involves adjustment of predictor scores in the direction of socially desirable outcomes which may help rectify inequities among groups. Novick and Peterson (1976) express reluctance over advocacy of a policy which is tantamount to reverse discrimination, but see the reverse discrimination problem as potentially alleviated <u>if</u> degree of disadvantage is used in adjusting test scores rather than racial or ethnic membership.

It is premature to speculate on which conception of test bias will prevail among scholars and practitioners (courts and federal guidelines?). Because the second and third conceptions of test bias imply continued use of test scores as part of the decision-making process, it appears not only appropriate but necessary to continue examinations of potential bias in test use. As Humphreys put it, "moves to abolish tests are more ostrich-like than human-like. The problem will simply not go away." (p. 66, 1972)

The present analysis was undertaken to examine potential bias in test use in the context of a widely used individual intelligence test and a popular standardized achievement test. Previous data on bias in test use were generally collected on college student samples or in adult selection-employment settings. The present study provides data on school age children using the Wechsler Intelligence Scale for Children - Revised (WISC-R) and the Metropolitan Achievement Test (MAT) from four ethnic groups and five grade levels.

Method

Sample and Procedure

In November, 1973 the Division of Special Education, Arizona State Department of Education funded a comprehensive study of handicapping conditions among school age children. The Pima County Special Services Cooperative was authorized to conduct a prevalence study within Pima County, a portion of which is reported herein. Pima County is geographically large (9200 square miles), ethnically diverse (approximately 68% Anglo, 25% Mexican American, 4% Black, and 3% American Indian), and largely urban in population (Tucson) with extensive and sparsely populated rural areas.

A stratified random sample of 1040 children was selected with equal numbers from four ethnic-racial groups (Anglo, Black, Mexican American, and Papago Indian with N = 260 per group), grade level (1st, 3rd, 5th, 7th, and 9th), sex, and urban-rural residence. The entire sample of Black children was selected from the urban area and the entire sample of Papago Indian children was selected from the rural area due to the very low proportion of urban Indians and rural Blacks in Pima County.

The cooperation of Tucson District #1, which enrolls about 2/3 of all school age children in the county, and all the rural school districts in the county was obtained through contacts with district authorities. School district enrollment rosters were used to randomly select the sample. Ethnicity was determined by school data, and in some cases, by contacting parents. Tucson District #1 was regarded as urban. Outlying districts, 25 miles or more from Tucson, were regarded as rural. Parents of children selected in the initial sample were contacted by letter or phone to explain the nature of the study and to solicit written permission. If parent permission was not obtained due to refusal, no reply, or the child withdrew from school, parents moved, etc., another child was selected from an alternative sample constituted by the above process. In the original sample of 1040 children, 67% of the parents granted permission, 4% refused permission, 18% did not reply and 11% could not be contacted due to absence of address and phone or because the family had moved. There were no appreciable differences among the groups in percentages of parents granting permission, refusing permission, no reply, or no address - family moved, etc.

As soon as parent permission was obtained, appointments were made with school officials to administer the various assessment procedures. The WISC-R was administered by appropriately trained examiners and all WISC-R protocols were further checked by the senior author for clerical and scoring errors. The MAT was usually administered in small groups under standard conditions.

Data Analysis

All data were analyzed by a procedure that uses the rationale presented by Gulliksen and Wilks (1950) and applied by Temp (1971). This procedure tests

the hypothesis that the regression systems for the groups are essentially the same. The homogeneity of errors of estimate, slopes, and intercepts of the separate regression equations are tested sequentially. Significant results at any stage of the analysis leads to rejection of the hypothesis of a common regression plane for the groups. In the present study the procedure was used to examine whether the relationship between the WISC-R and MAT was the same for all groups.

Results

Due to various logistical problems, e.g., delays in receipt of parental permissions, school scheduling problems, and availability of examiners to travel to remote areas, WISC-R scores were obtained for only 950 of the original sample of 1040 students. Of the 950 students for whom WISC-R scores were available, MAT scores were obtained for 910 students (Anglo N = 250, Black N = 222, Mexican-American N = 215, and Papago Indian N = 223). The results are based on the 910 students for whom both scores were available.

The results of the Gulliksen-Wilks testing procedure to assess whether the regression systems are equal within each grade and pair of tests are presented in Table 1. The standard errors of estimate are tested first by this procedure, and as indicated in Table 1, 14 of the 30 tests resulted in rejection of the hypothesis of equal errors of estimate ($p \leq .05$). The slopes are then examined in this procedure, and 3 of the remaining 16 sets regression lines were found to have unequal slopes. The 13 remaining sets of regression equations were then examined for equality of intercepts, and 10 more were rejected. Overall, out of a total of 30 sets of regression equations, the hypothesis of a common regression system for the four groups was retained in only 3 cases.

In Table 2 the data on standard errors of estimate, slopes, and intercepts for the four groups and five grade levels are presented.

Discussion and Summary

A search of the literature revealed three somewhat contradictory conceptualizations of test bias. The first definition which requires equality of means among all groups was not analyzed with the present data. We are quite skeptical about the value of this definition in that it seems to imply that measures of performance are acceptable only if everyone performs at the same level. Furthermore, the remedies which have sometimes been suggested by proponents of this definition, i.e., nonstandard-administration of tests and/or development of separate tests and/or norms for specific racial or ethnic groups, are seen as counterproductive in overcoming the persistent problems of social class, ethnic, and racial discrimination in society.

The second conceptualization of bias in test use requires equivalent prediction systems for all groups. Contrary to trends in previous studies (see Horn, 1976; or Humphreys, 1972 for reviews), our data indicate that the relationship between two widely used standardized tests differs substantially among four groups. The differences in the outcome of this study of test bias and previous studies may be related to differences in age level or ethnic or racial groups studied, or differences in method used to examine the equality of the regression systems. Most previous studies have examined only differences in slope or differences in slope and intercept. In the present study tests of equality of standard errors of estimate were also conducted, and resulted in a large number of rejection decisions.

It is perhaps important to note that our present analysis does not indicate direction of bias, e.g., whether errors of estimate are systematically smaller or larger for specific groups, whether over or underprediction results for specific groups if a common regression line is used, etc. These data do indicate that the assumption of a common regression system for the two tests and four ethnic groups studied is largely untenable.

Finally, our data are only partially relevant to the third conceptualization of test bias. Although we strongly resist the use of tests to determine social policy, we recognize the very legitimate concerns about the social consequences of test use. The social consequences of test use, and attempts to rectify current social inequities through adjustment of test scores, are issues which cannot be resolved within the realm of strictly empirical approaches. Along with Novick and Peterson (1976) we have reservations about the use of ethnicity or race to mechanically adjust test scores and/or prediction equations. The recent work of Mercer and Lewis (1976) in developing the System of Multicultural Pluralistic Assessment (SOMPA) provides one model for incorporating social background and ethnic or racial data in test interpretations. Further research and broadly based social policy discussions are needed regarding these alternative suggestions for test use.

References

Boehm, V. Negro-white differences in validity of employment and training selection procedures: Summary of recent evidence. Journal of Applied Psychology, 1972, 56, 33-39.

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

Cole, N. Bias in selection. Journal of Educational Measurement, 1973, 10, 237-255.

Darlington, R. Another look at "cultural fairness." Journal of Educational Measurement, 1971, 8, 71-82.

Gulliksen, H., & Wilks, S. Regression tests for several samples. Psychometrica, 1950, 15, 91-114.

Horn, J. Human abilities: A review of research and theory in the early 1970s. p. 437-485, in Annual Review of Psychology, Vol. 27, 1976.

Humphreys, L. Implications of group differences for test interpretation. Proceedings of 1972 Invitational Conference on Testing Problems, ETS, Princeton, New Jersey, 1973.

Jackson, G. On the report of the ad hoc committee on educational uses of tests with disadvantaged students. American Psychologist, 1975, 30, 88-92.

Kallingal, A. The prediction of grades for black and white students at Michigan State University. Journal of Educational Measurement, 1971, 8, 263-265.

Larry P. et al. v. Wilson Riles, et al. United States District Court, Northern District of California, Case No. C-71-2270 RFP.

Mercer, J., & Lewis, J. System of Multicultural Pluralistic Assessment. Psychological Corporation, 1976, in press.

Novick, M., & Peterson, N. Towards equalizing educational and employment opportunity. Journal of Educational Measurement, 1976, 13, 77-88.

Peterson, N., & Novick, M. An evaluation of some models for culture fair selection. Journal of Educational Measurement, 1976, 13, 3-29.

Pfiefer, C., & Sedlacek, W. The validity of academic predictors for black and white students at a predominantly white university. Journal of Educational Measurement, 1971, 8, 253-261.

Schmidt, F., Berner, J., & Hunter, J. Racial differences in validity of employment tests. Journal of Applied Psychology, 1973, 58, 5-9.

Temp, G. Validity of the SAT for blacks and whites in thirteen integrated institutions. Journal of Educational Measurement, 1971, 8, 245-251.

Thorndike, R. Concepts of culture fairness. Journal of Educational Measurement, 1971, 8, 63-70.

Williams, R. The problem of the match and mismatch in testing black children. In L. Miller (Ed.), The Testing of Black Students: A Symposium. Englewood Cliffs, N.J.: Prentice Hall, 1974.

Table 1

Summary of Gulliksen-Wilks Tests of Equality of Regression Lines[1]

		Error of estimate	Slope	Intercept	Single regression plane
First Grade	MAT Reading on Verbal IQ	Yes	-	-	No
	MAT Reading on Perf. IQ	Yes	-	-	No
	MAT Reading on Full Scale	Yes	-	-	No
	MAT Math on Verbal IQ	N:S	N:S	Yes	No
	MAT Math on Perf. IQ	N:S	N:S	Yes	No
	MAT Math on Full Scale	N:S	N:S	Yes	No
Third Grade	MAT Reading on Verbal IQ	Yes	-	-	No
	MAT Reading on Perf. IQ	Yes	-	-	No
	MAT Reading on Full Scale	Yes	-	-	No
	MAT Math on Verbal IQ	N:S	N:S	Yes	No
	MAT Math on Perf. IQ	N:S	N:S	Yes	No
	MAT Math on Full Scale	N:S	N:S	Yes	No
Fifth Grade	MAT Reading on Verbal IQ	N:S	Yes	-	No
	MAT Reading on Perf. IQ	Yes	-	-	No
	MAT Reading on Full Scale	N:S	Yes	-	No
	MAT Math on Verbal IQ	N:S	N:S	N:S	Yes
	MAT Math on Perf. IQ	Yes	-	-	No
	MAT Math on Full Scale	N:S	N:S	Yes	No
Seventh Grade	MAT Reading on Verbal IQ	N:S	N:S	N:S	Yes
	MAT Reading on Perf. IQ	N:S	Yes	-	No
	MAT Reading on Full Scale	N:S	N:S	N:S	Yes
	MAT Math on Verbal IQ	Yes	-	-	No
	MAT Math on Perf. IQ	Yes	-	-	No
	MAT Math on Full Scale	Yes	-	-	No
Ninth Grade	MAT Reading on Verbal IQ	N:S	N:S	Yes	No
	MAT Reading on Perf. IQ	N:S	N:S	Yes	No
	MAT Reading on Full Scale	N:S	N:S	Yes	No
	MAT Math on Verbal IQ	Yes	-	-	No
	MAT Math on Perf. IQ	Yes	-	-	No
	MAT Math on Full Scale	Yes	-	-	No

[1] All tests were conducted at the .05 level.

N:S indicates nonsignificant differences.

All tests are sequential. A significant outcome indicated in table by Yes indicates that no further tests should be conducted.

Table 2

Standard errors of Estimate, Slope, Intercept and Correlations for MAT and WISC-R by Group and Grade

GRADE 1		SEest	Slope	Intercept	r
	Reading by Verbal IQ				
N = 49	Anglo	9.32	.070	49.2	.13
N = 40	Black	6.52	.361	16.0	.65
N = 44	Mexican-American	9.08	.319	25.6	.44
N = 48	Papago Indian	7.06	.248	25.0	.36
	Reading by Performance IQ				
	Anglo	9.34	.076	48.3	.11
	Black	6.70	.471	5.8	.63
	Mexican-American	9.14	.337	20.2	.43
	Papago Indian	6.74	.265	20.7	.46
	Reading by Full Scale IQ				
	Anglo	9.28	.110	45.2	.16
	Black	6.09	.448	8.5	.70
	Mexican-American	8.80	.406	16.2	.49
	Papago Indian	6.63	.347	15.9	.48
	Math by Verbal IQ				
	Anglo	8.32	.187	38.1	.37
	Black	7.40	.347	16.3	.59
	Mexican-American	7.88	.421	16.3	.60
	Papago Indian	7.30	.333	19.8	.45
	Math by Performance IQ				
	Anglo	8.80	.112	44.9	.18
	Black	7.71	.429	8.6	.53
	Mexican-American	9.35	.237	29.4	.31
	Papago Indian	7.06	.315	17.5	.50
	Math by Full Scale IQ				
	Anglo	8.39	.222	34.2	.34
	Black	7.25	.411	10.7	.61
	Mexican-American	8.35	.424	14.1	.53
	Papago Indian	6.73	.439	9.8	.57

Table 2 - Continued

GRADE 3		SZest	Slope	Intercept	r
	Reading by Verbal IQ				
N = 51	Anglo	8.74	.457	11.0	.55
N = 40	Black	6.15	.394	17.7	.69
N = 45	Mexican-American	7.73	.298	24.4	.42
N = 51	Papago Indian	4.88	.225	26.7	.48
	Reading by Performance IQ				
	Anglo	8.68	.405	15.6	.56
	Black	8.09	.285	27.1	.31
	Mexican-American	7.39	.362	17.3	.50
	Papago Indian	4.87	.248	20.1	.48
	Reading by Full Scale IQ				
	Anglo	8.22	.492	7.4	.62
	Black	6.53	.484	10.2	.64
	Mexican-American	7.21	.415	14.0	.54
	Papago Indian	4.68	.290	18.5	.54
	Math by Verbal IQ				
	Anglo	8.34	.357	20.6	.47
	Black	7.64	.186	33.1	.34
	Mexican-American	9.72	.177	36.1	.21
	Papago Indian	7.03	.147	31.8	.24
	Math by Performance IQ				
	Anglo	8.31	.315	27.3	.48
	Black	7.70	.283	24.4	.32
	Mexican-American	9.09	.344	20.1	.41
	Papago Indian	6.81	.227	23.1	.34
	Math by Full Scale IQ				
	Anglo	8.00	.384	17.8	.53
	Black	7.52	.274	25.7	.38
	Mexican-American	9.35	.311	24.1	.34
	Papago Indian	6.87	.221	25.3	.32

Table 2 - Continued

GRADE 5		SEest	Slope	Intercept	r
	Reading by Verbal IQ				
N = 52	Anglo	6.36	.514	6.0	.80
N = 45	Black	5.88	.448	11.0	.72
N = 48	Mexican-American	5.96	.430	12.8	.70
N = 44	Papago Indian	6.31	.123	36.3	.20
	Reading by Performance IQ				
	Anglo	9.74	.380	20.0	.39
	Black	8.04	.180	31.8	.30
	Mexican-American	7.68	.256	25.2	.40
	Papago Indian	6.31	.110	35.5	.19
	Reading by Full Scale IQ				
	Anglo	6.79	.628	-5.4	.77
	Black	6.90	.373	16.6	.57
	Mexican-American	6.44	.424	12.2	.64
	Papago Indian	6.29	.137	34.5	.21
	Math by Verbal IQ				
	Anglo	7.53	.424	14.0	.68
	Black	6.98	.452	11.7	.66
	Mexican-American	6.82	.302	25.8	.52
	Papago Indian	5.07	.347	18.2	.58
	Math by Performance IQ				
	Anglo	9.57	.341	22.7	.36
	Black	8.14	.315	21.0	.47
	Mexican-American	7.50	.207	32.0	.34
	Papago Indian	6.10	.106	33.6	.19
	Math by Full Scale IQ				
	Anglo	7.75	.521	4.3	.65
	Black	7.09	.460	10.4	.64
	Mexican-American	6.97	.307	24.6	.49
	Papago Indian	5.59	.274	21.9	.44

Table 2 - Continued

GRADE 7		SEest	Slope	Intercept	r
	Reading by Verbal IQ				
N = 54	Anglo	7.09	.472	9.9	.63
N = 51	Black	6.69	.515	3.4	.75
N = 46	Mexican-American	7.10	.388	15.6	.52
N = 43	Papago Indian	6.16	.347	18.7	.53
	Reading by Performance IQ				
	Anglo	8.43	.227	34.0	.38
	Black	7.62	.532	1.3	.65
	Mexican-American	8.13	.144	35.4	.20
	Papago Indian	6.84	.180	27.5	.34
	Reading by Full Scale IQ				
	Anglo	7.55	.776	19.1	.56
	Black	6.42	.582	-2.1	.77
	Mexican-American	7.38	.388	14.5	.46
	Papago Indian	6.43	.289	20.9	.47
	Math by Verbal IQ				
	Anglo	8.67	.461	9.4	.54
	Black	7.72	.391	15.8	.59
	Mexican-American	7.09	.491	7.9	.61
	Papago Indian	5.07	.161	30.8	.33
	Math by Performance IQ				
	Anglo	9.39	.281	26.9	.41
	Black	8.46	.366	17.5	.47
	Mexican-American	8.65	.194	31.9	.26
	Papago Indian	5.02	.141	29.8	.36
	Math by Full Scale IQ				
	Anglo	8.77	.401	15.0	.53
	Black	7.72	.427	12.8	.59
	Mexican-American	7.50	.496	6.1	.55
	Papago Indian	4.95	.179	28.3	.39

Table 2 - Continued

GRADE 9		SEest	Slope	Intercept	r
	Reading by Verbal IQ				
N=44	Anglo	6.12	.470	9.8	.72
N=46	Black	7.30	.353	16.7	.55
N=32	Mexican-American	6.80	.432	12.7	.62
N=37	Papago Indian	5.99	.321	19.6	.49
	Reading by Performance IQ				
	Anglo	7.80	.365	20.8	.46
	Black	8.40	.166	32.8	.28
	Mexican-American	7.53	.327	19.2	.50
	Papago Indian	6.52	.167	28.1	.31
	Reading by Full Scale IQ				
	Anglo	6.25	.519	4.4	.70
	Black	7.76	.294	22.1	.46
	Mexican-American	6.77	.449	10.0	.63
	Papago Indian	6.04	.327	16.8	.47
	Math by Verbal IQ				
	Anglo	8.65	.512	5.6	.62
	Black	6.00	.182	30.9	.38
	Mexican-American	7.05	.477	8.7	.65
	Papago Indian	4.79	.220	27.9	.43
	Math by Performance IQ				
	Anglo	9.55	.503	6.6	.50
	Black	6.34	.095	38.5	.21
	Mexican-American	7.80	.373	14.7	.54
	Papago Indian	5.26	.054	39.3	.13
	Math by Full Scale IQ				
	Anglo	8.32	.612	-5.2	.66
	Black	6.14	.155	33.4	.33
	Mexican-American	6.88	.509	4.5	.67
	Papago Indian	5.02	.174	30.1	.32