DOCUMENT RESUME

ED 222 529 TM 820 677

AUTHOR Reschly, Daniel J.; Ross-Reynolds, Jane

TITLE An Investigation of WISC-R Item Bias with Black, Chicano, Native American Papago, and White Children: Implications for Nondiscriminatory Assessment.

PUB DATE 82

NOTE 20p.

EDRS PRICE MF01/PC01 Plus Postage.

DESCRIPTORS Academic Ability; Correlation; Difficulty Level; *Exceptional Persons; *Intelligence Tests; *Minority Group Children; Psychometrics; Racial Differences; Student Evaluation; *Test Bias; *Test Items; Test Reliability

IDENTIFIERS *Nondiscriminatory Assessment; Outliers; *Wechsler Intelligence Scale for Children (Revised)

ABSTRACT

Two psychometric approaches for the investigation of item bias were employed to examine the performance of four sociocultural groups on six subtests of the Wechsler Intelligence Scale for Children - Revised (WISC-R). The sample of 950 students was from Pima County, Arizona. One approach involved the comparison of the psychometric indices of internal consistency reliability, which were reasonably high and comparable for the four groups; the rank order correlations of item difficulty, which were quite high and comparable for the four groups; and the rank order correlations of the differences in the difficulty of adjacent items, which were less comparable for Native American Papagos suggesting possible items by group interactions. A second approach involved the utilization of transformed item difficulties and a quantitative method of outlier analysis to identify specific items as biased (i.e., as manifesting an item by group interaction). Results were not interpreted as evidence of cultural bias, however, since the possible confounding effects of socioeconomic status and overall ability were not controlled. Further, the items identified did not appear to reflect aspects of the assumption of "interference" implicit in the notion of cultural bias. (Author/PN)

AN INVESTIGATION OF WISC-R ITEM BIAS WITH BLACK, CHICANO, NATIVE AMERICAN PAPAGO, AND WHITE CHILDREN: IMPLICATIONS FOR NONDISCRIMINATORY ASSESSMENT

Daniel J. Reschly
Department of Psychology
Iowa State University
Ames, Iowa 50011

and

Jane Ross-Reynolds
Terrebonne Parish Schools
Thibodaux, LA 70301

ABSTRACT

Two psychometric approaches for the investigation of item bias were employed to examine the performance of four sociocultural groups on six subtests of the Wechsler Intelligence Scale for Children - Revised (WISC-R). The sample of 950 students (Anglo N = 252; Black N = 237; Chicano N = 223; and Native American Papago N = 238) was from Pima County, Arizona. The subtests analyzed were the five Verbal Scale subtests of Information, Similarities, Arithmetic, Vocabulary, and Comprehension and the Performance Scale subtest of Picture Completion. One approach involved the comparison of the psychometric indices of internal consistency reliability which were reasonably high and comparable for the four groups, the rank order correlations of item difficulty which were quite high and comparable for the four groups, and the rank order correlations of the differences in the difficulty of adjacent items which were less comparable for Native American Papagos suggesting possible items by group interactions. A second approach involved the utilization of transformed item difficulties and a quantitative method of outlier analysis to identify specific items as biased (i.e., as manifesting an item by group interaction). No items were identified as biased against Blacks on any of the subtests. Only one item was identified as biased against Chicanos. No items were identified as biased against Native American Papagos on the Performance subtest, Picture Completion, but nearly one-third of the items on the Verbal subtests were so identified. A majority of the items on the six subtests were slightly more difficult for the Black, Chicano, and Native American Papago groups than for the Anglo group. These results were not interpreted as evidence of cultural bias, however, since the possible confounding effects of socioeconomic status and overall ability were not controlled. Further, the items identified did not appear to reflect aspects of the assumption of "interference" implicit in the notion of cultural bias.

INTRODUCTION

Allegations of cultural bias against Black, Hispanic, and Native American children in individually administered intelligence tests have been made both in the academic literature (Gay & Abrahams, 1973; Laosa, 1977; Mercer & Brown, 1973; Williams, 1971, 1975) and by plaintiffs in court cases (Diana v. California State Board of Education, 1970; Guadalupe v. Tempe School District, 1971; Larry P. v. Riles, 1972, 1974, 1979; PASE v. Hannon, 1980). Those alleging cultural bias have often done so on the basis of their judgment that the domain sampled by intelligence tests has been restricted to the experiences of white, middle class groups. "American IQ tests have, inevitably, included items and procedures which reflect the abilities and skills valued by the American core culture. This 'core culture' consists mainly of the cultural patterns of that segment of the population consisting of white, Anglo-Saxon Protestants whose social status today has become middle and upper middle class" (Mercer & Brown, 1973, p. 66).

A variety of empirical investigations of cultural bias have been made using different definitions of bias, different tests, and different populations. The majority have focused upon tests used for the selection of candidates for college admission or employment with adolescent and adult populations. Studies using individual intelligence tests with children have by and large been investigations of atmosphere bias (Sattler, 1974). Substantially fewer studies have been made of item bias, especially on individual intelligence tests with this age group.

Empirical investigations of item bias on the Wechsler Intelligence Scale for Children - Revised (WISC-R) (Wechsler, 1974) are needed on a number of grounds. First, as Judge Peckham observed in his Opinion (Larry P. v. Riles, 1979), there is a lack of information upon which to base a decision regarding this test. Secondly, there has been a tendency among those alleging cultural bias to select specific items for criticism in order to indict the entire test. For example, Item 6 on the WISC-R Comprehension subtest ("What is the thing to do if a boy/girl much smaller than yourself starts to fight with you?") has often been cited as biased against black children (APA Monitor, 1977; Williams, 1975). As a matter of fact, Judge Peckham relied upon this criticism in his Opinion. "Cultural differences can also be found in specific test items. Some of these items have in fact become rather notorious, such as the 'fight item' on WISC tests. This question asked children what they would do if struck by a smaller child of the same sex. The 'correct' answer is that it is wrong to strike the child back... Similarly, it may be that such questions as who wrote Romeo and Juliet, who discovered America, and who invented the light bulb, are culturally biased" (Opinion, Larry P. v. Riles, 1979, p. 48).

Two approaches have principally been employed in empirical research on item bias. One approach, elaborated by Arthur Jensen (1978; 1980), involves the comparison of selected psychometric features of a test such as the internal consistency reliability, the item correlation with total score, the rank order of item difficulties, and the relative difficulty of adjacent items. These psychometric characteristics would be expected to be similar of the test functioned the same for the groups under comparison. Bias would be suspected if these psychometric properties were different for the groups in question. The second approach, associated primarily with the process of test construction, includes a variety of methods that have been developed for identifying specific items as biased. One method incorporates subjective judgment by having panels of experts scrutinize the items to detect biased content. Unfortunately, the degree of agreement among the experts has tended to be very low (Flaugher, 1978; Jensen, 1977). Among the other methods of analyzing item bias, the principal ones may be designated as: transformed item difficulty (Anastasi, 1976); correlational (Green & Draper, 1972 cited in Merz & Rudner, Note 1); chi square (Scheuneman, 1979); item characteristic curve (Lord, 1977 and others cited in Merz & Rudner, Note 1); and distractor response analysis (Veale & Foreman, 1975, 1976 cited in Merz & Rudner, Note 1). An analysis of variance framework has also been used to investigate bias defined as an item by group interaction with test items and group membership considered main effects (Angoff & Ford, 1973; Cleary & Hilton, 1968). This method has not permitted the identification of specific items as biased, however, unless accompanied by numerous a priori comparisons or post hoc tests (Merz & Rudner, Note 1).

These methods have typically been employed with large samples and group-administered tests on which it may be assumed that all examinees have had an opportunity to respond to all items. Of those mentioned, the transformed item difficulty method is probably the oldest and best known (Anastasi, 1976).

According to the transformed item difficulty method, a specific item is considered biased if it appears relatively more difficult for one group than another, i.e., if there is an item by group interaction visible on a scatterplot. Merz and Rudner (Note 1) suggest that biased items be identified quantitatively by means of an outlier analysis. One method of outlier analysis that they describe involves setting a fixed boundary of .75 z-score units. Outlying items beyond this boundary are considered biased.

Only one study of item bias on the WISC-R has been published previously (Sandoval, 1979). Analyzing WISC-R data obtained in the standardization of the System of Multicultural Pluralistic Assessment (Mercer, 1979), Sandoval reported that: 1) the Cronbach Alpha coefficients indicated high and comparable internal consistency reliabilities for the Anglo, Black, and Chicano groups on all of the subtests; 2) using item means as a measure of item difficulty, the rank order correlations of item difficulties and of the differences in adjacent item difficulties were all quite high; 3) the amount of variance attributable to the item by ethnic group interaction (an indicator of bias) and to the item by SES within ethnic group interaction was generally small, but statistically significant, particularly on the Verbal subtests; and 4) a total of 97 items were identified as most discriminatory between the Anglo and Black or Chicano groups in a Multivariate Analysis of Variance. Sandoval concluded that "In general the notion that there may be a number of items with radically different difficulties for children from different ethnic groups has not been supported... The lack of a clear pattern of difficult items and the fact that there exist a large number of items just slightly more difficult for minority group children spread throughout the entire test suggests that general factors rather than specific item content contribute to differences in means" (Sandoval, 1979, p. 925).

The purpose of this study, like the one by Sandoval, was the empirical investigation of item bias on the WISC-R. The present study was different in that the performance of a group of Native American children was examined as well as the performance of groups of Anglo, Black, and Chicano children. In addition, the method varied in that transformed item difficulties and a method of outlier analysis for the identification of biased items were applied to the data. However, some of the analyses were similar in the two studies permitting a comparison of the results.

The basic design of this study incorporated methods representative of the two principal approaches to the investigation of item bias reported in the literature. Accordingly, two major questions were addressed. First, were such psychometric properties as the internal consistency reliability, the rank order of item difficulty, the differences in the difficulties of adjacent items, and the correlations between item performance and subtest score similar for the four groups? Second, using transformed item difficulties and a method of outlier analysis for the identification of biased items, which and how many items were so identified?

METHOD

The data were collected in the Pima County Prevalence Study which has been described in detail by Reschly and Jipson (1976). A random sample of 1,040 children was selected from enrollment rosters of all schools in Pima County, Arizona. Equal representation was given to ethnic group (Anglo, Black, Chicano, and Native American Papago N = 260), grade level (1st, 3rd, 5th, 7th, and 9th), and sex. The final sample sizes were 252 Anglos; 237 Blacks; 223 Chicanos; and 238 Native American Papagos. The scores of all twelve subtests of the WISC-R were obtained for each child. The WISC-R was administered by appropriately trained examiners, and all protocols were checked for clerical and scoring errors.

The Wechsler Intelligence Scale for Children - Revised was published in 1974 (Wechsler, 1974). The instrument has been extensively reviewed by Sattler (1982). The administration and scoring procedures are clearly detailed in the Manual (Wechsler, 1974). The analyses which follow had to take into account two particular facets of the administration and scoring of the WISC-R. Specifically, all subjects did not receive all items, and responses were not scored on a simple right/wrong, 1 or 0, basis. As a consequence, the only Performance subtest included was Picture Completion. The p-values (percentage passing each item) were obtained by dividing the number passing the item by the total number of subjects since the subjects would not have received credit for items beyond their discontinue points according to standardized scoring procedures. Finally, for the Verbal subtests, Similarities, Vocabulary, and Comprehension, passing was arbitrarily defined as a score of 1 or more.

The data were analyzed as follows:

1: The raw score means and standard deviations for each group were calculated for each subtest and for the Verbal and Performance scales (with Mazes substituted for Coding).

2. Internal reliability coefficients were calculated using Cronbach Alpha for each of the four groups for 10 of the 12 subtests (Coding and Digit Span not included).

3. For the subtests, Information, Similarities, Arithmetic, Vocabulary, Comprehension, and Picture Completion, p-values were calculated for each of the four groups, Anglos, Blacks, Chicanos, and Native American Papagos. The items on each subtest were assigned ranks according to p-values. Rank order correlations (Spearman Rho) of p-values between the four groups were then calculated.

4. Differences in the p-values of adjacent items for each group were computed for the six subtests, and rank order correlations (Spearman Rho) of these differences were then calculated.

5. According to the transformed item difficulty method, p-values were transformed to *z*-values and then to delta values using a linear transformation ($4z + 21$) to eliminate the negative signs. A low delta value thus indicated a relatively easy item. A high delta value indicated a more difficult item. Delta values ranked from 1 (100% passing) to 41 (0% passing). Given the administration and scoring rules of the WISC-R in which subjects received credit for items below their entry point and did not receive credit for items beyond their discontinue point, the items with average values for the four groups corresponding to more than 95% passing and fewer than 5% passing were arbitrarily eliminated. These items were considered subject to basal and ceiling effects.

6. Pairs of delta values for each of the remaining items were then plotted on graphs. Three scatterplots, comparing the delta values of the Anglo group with those of the Black, Chicano, and Native American Papago groups, were prepared for each subtest. A 45° line was drawn through the origin on each scatterplot to represent a theoretical regression line of equal item difficulties for the two groups.

7. Biased items were defined as those exhibiting an item by group interaction and were identified by means of an outlier analysis. The items with average *z*-values for the four groups corresponding to more than 95% passing and fewer than 5% passing were not included in this analysis since these items typically were not administered to the vast majority of participants. The distance of each item point to the 45° theoretical regression line was calculated using the equation $d = (z_1 - z_2) / \sqrt{2}$ in which z_j was the *z*-value of the item difficulty for group j. Outlying items at a distance greater than .75 were categorized as biased (Merz & Rudner, Note 1). The Anglo group was taken as the basis for comparison since the WISC-R has been characterized as reflecting Anglo culture.

8. The point biserial correlations between item performance and raw subtest score were determined for the subtests of Information, Arithmetic, and Picture Completion for each group. The overall pattern of the correla-

tions was compared, and the character of the correlations for the items identified by the previous analysis as biased was considered. Items with point biserial correlations above .40 were judged to be discriminating well.

RESULTS

Means and Standard Deviations

Table 1 presents the means and standard deviations of the raw scores for the four sociocultural groups on all subtests and the Verbal and Performance Scales. The standard deviations on all subtests were quite comparable, though generally lower for Blacks, Chicanos, and Papagos on the Verbal subtests. However, the differences in the standard deviations among the four groups were slight, except on the Vocabulary subtest. The standard deviation for the Anglo group on Vocabulary was one-third larger than the standard deviations of the other three groups. On all of the subtests, the means were lower for the Black, Chicano, and Papago groups by at least one-half of the standard deviations of the Anglo group.

Table 1

Raw Score Means and Standard Deviations

	Anglos		Blacks		Chicanos		Papagos	
	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
Information	14.76	5.74	11.43	4.76	11.25	4.89	9.76	3.84
Similarities	13.74	5.91	10.47	5.41	10.00	5.33	7.32	4.51
Arithmetic	11.60	3.32	10.14	3.37	10.17	3.07	8.61	3.00
Vocabulary	31.40	12.10	26.11	9.67	24.18	9.10	19.29	8.19
Comprehension	18.65	7.60	15.27	6.80	14.79	6.50	10.57	5.69
Picture Completion	17.77	4.91	15.53	5.01	16.56	4.58	15.59	4.39
Picture Arrangement	26.45	9.52	20.95	10.72	21.90	9.75	20.61	10.52
Block Design	27.90	14.98	19.46	12.43	22.57	13.91	21.56	14.19
Object Assembly	21.34	5.87	17.93	5.48	18.90	5.98	17.82	6.17
Mazes	21.14	5.78	18.18	6.70	20.40	5.96	20.70	5.82
Verbal Scale	90.15	32.46	73.43	27.76	70.40	26.43	55.55	22.39
Performance Scale[a]	114.60	35.60	92.04	34.12	100.32	33.86	96.27	34.77
	N=252		N=237		N=223		N=238	

[a] The Performance Scale was calculated by substituting Mazes for Coding.

Internal Consistencies

As shown in Table 2, the Cronbach Alpha reliability coefficients were very comparable for the four groups, and, with one exception (Blacks on Block Design) were higher than the Spearman-Brown coefficients reported in the Manual (Wechsler, 1974). On the basis of this analysis alone, it is not possible to state whether the coefficients are truly different from one another either among the four groups or between the two types of estimates of internal reliability. One may observe, however, that the test has reasonably high and comparable reliability for the four groups; and the reliability coefficients reported in the Manual would appear to be appropriate for computing the standard error of measurement to be used for all four groups.

Table 2

Cronbach Alpha Coefficients

	Anglos	Blacks	Chicanos	Papagos
Information	.913	.893	.895	.865
Similarities	.858	.859	.856	.825
Arithmetic	.848	.862	.839	.842
Vocabulary	.927	.904	.893	.882
Comprehension	.900	.891	.882	.848
Picture Completion	.892	.892	.866	.852
Picture Arrangement	.786	.840	.784	.817
Block Design	.873	.842	.864	.871
Object Assembly	.797	.711	.748	.737
Mazes	.722	.796	.743	.703

Average reliability coefficients across age groups presented in manual: calculated according to Spearman-Brown formula for split-half reliability

Information	.85
Similarities	.81
Arithmetic	.77
Vocabulary	.86
Comprehension	.77
Picture Completion	.77
Picture Arrangement	.73
Block Design	.85
Object Assembly	.70
Mazes	.72

Rank Order Correlations

The rank order correlations (Spearman Rho) of p-values displayed in Table 3 reveal the degree to which the ordering of item difficulties was similar for the four groups on the subtests of Information, Similarities, Arithmetic, Vocabulary, Comprehension, and Picture Completion. The correlations were all quite high, the lowest being .97 between Anglos and Blacks on Comprehension.

Table 3

Rank Order Correlations of p-values

	Anglos x Blacks	Anglos x Chicanos	Anglos x Papagos	Blacks x Chicanos	Blacks x Papagos	Chicanos x Papagos
Information	.983	.988	.987	.992	.984	.988
Similarities[a]	.985	.988	.995	.990	.988	.992
Arithmetic	.991	.991	.985	.996	.991	.992
Vocabulary[a]	.992	.988	.989	.986	.983	.987
Comprehension[a]	.970	.976	.973	.982	.986	.994
Picture Completion	.991	.987	.978	.991	.986	.990

[a] A score of 2 or 1 was considered a passing score.

While the p-values appeared to be in much the same rank order for the four groups, the differences in the p-values of adjacent items did not. The rank order correlation of differences in the p-values of adjacent items given in Table 4 were lower than the rank order correlations of p-values presented earlier. These correlations suggest that the progression of difficulty was not quite as smooth as the rank order correlations of p-values would imply, particularly for Papagos, and that there were possible item by group interactions.

Table 4

Rank Order Correlations of Differences in Adjacent p-values

	Anglos x Blacks	Anglos x Chicanos	Anglos x Papagos	Blacks x Chicanos	Blacks x Papagos	Chicanos x Papagos
Information	.41	.66	.49	.74	.48	.75
Similarities[a]	.74	.68	.55	.69	.65	.86
Arithmetic	.84	.78	.57	.81	.69	.55
Vocabulary[a]	.57	.68	.39	.56	.58	.44
Comprehension[a]	.74	.85	.51	.68	.67	.74
Picture Completion	.69	.46	.21	.80	.53	.48

[a] A score of 2 or 1 was considered a passing score.

Transformed Item Difficulties

The item difficulties (p-values) transformed to delta values may be displayed on scatterplots. The scatterplots for information are illustrative (see Figure 1). Each point represents an item. The delta values of the items for Anglos are given on the abscissas, and for Blacks, Chicanos, and Papagos on the ordinates. A line drawn through the origin at a 45 angle represents points of equal item difficulty for the groups pictured.

The points representing the item difficulties for the groups on Information occur at or above the equal time difficulty line signifying that all of the items were relatively more difficult for Blacks, Chicanos, and Papagos than for Anglos. Nevertheless, straight lines could be fitted to the points to express linear relationships which would be consistent with the high rank order correlations of p-values. Table 5 lists the items with average delta values for the four groups correponding to p-values of more than 95% passing and fewer than 5% passing which were eliminated from further consideration since they were considered subject to basal and ceiling effects and, in fact, were not administered to most participants.

Table 5

Items Eliminated[a]

Information	Items 1, 2, 3, 25, 26, 27, 28, 29, 30
Similarities	Items 16, 17
Arithmetic	Items 1, 2, 3, 5, 18
Vocabulary	Items 1, 2, 3, 4, 5, 6, 25, 28, 29, 30, 31, 32
Comprehension	None
Picture Completion	Items 1, 2, 3, 4, 26

[a] Items with delta values corresponding to more than 95% passing or fewer than 5% passing

Outlier Analysis

Table 6 summarizes the results obtained by employing a method of outlier

Table 6

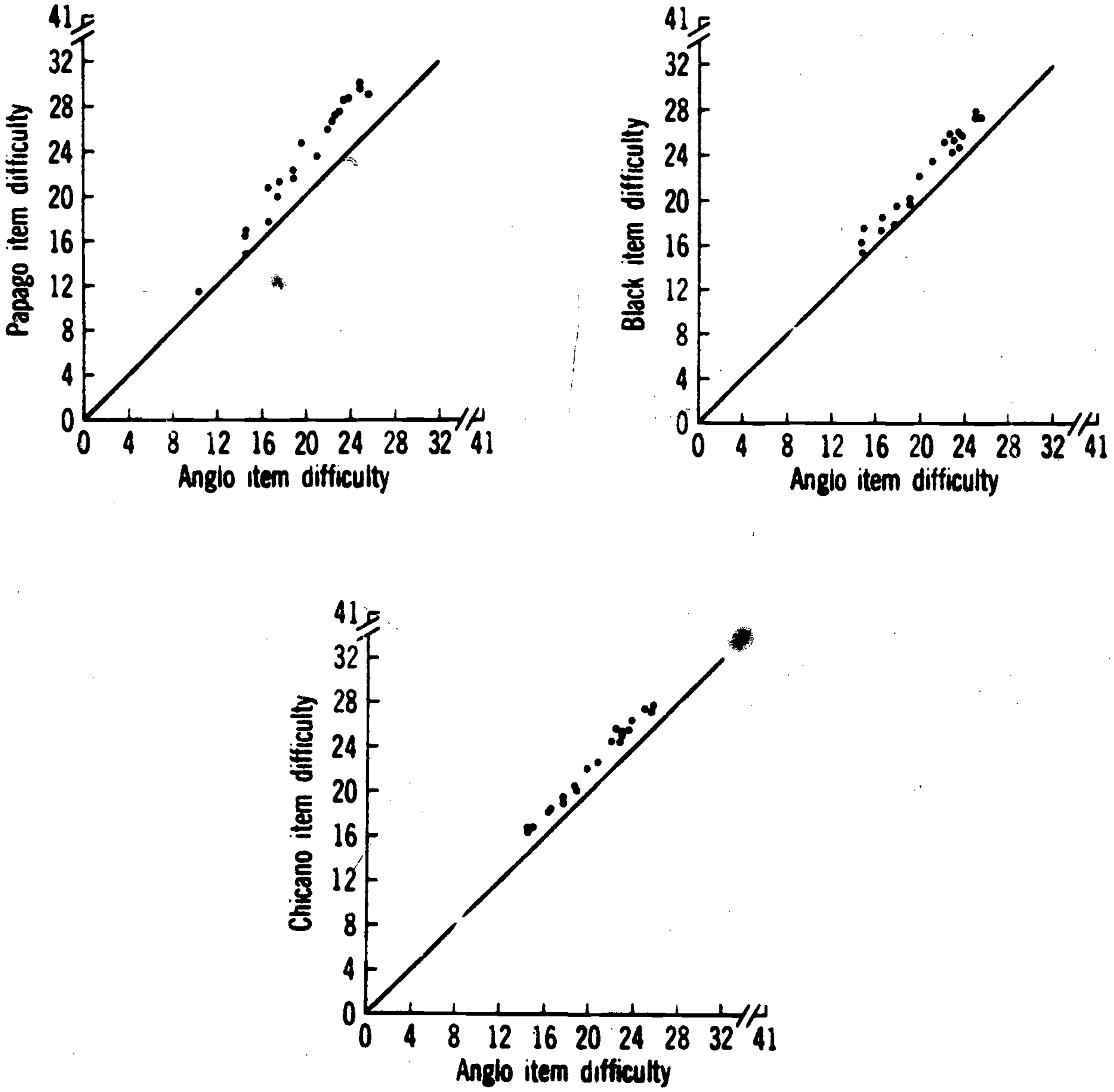
Outlying Items[a]

	Anglos x Blacks	Anglos x Chicanos	Anglos x Papagos
Information	None	None	Items 7, 13, 15, 16, 17, 18, 19, 20, 21, 22, 24
Similarities	None	None	Items 10, 11, 12, 13, 14
Arithmetic	None	None	Items 6, 13, 15, 17
Vocabulary	None	Item 22	Items 15, 16, 17, 18, 19, 22, 24, 27
Comprehension	None	None	Items 10, 11, 12, 13, 14, 16, 17
Picture Completion	None	None	None

[a] Items at a distance greater than .75 where distance equals $(z_1 - z_2)/\sqrt{2}$

Figure 1

Scatterplots of Transformed Item Difficulties on the Information Subtest



analysis. The outlying items were ones which were more difficult by a fixed amount for the Black, Chicano, or Papago groups than for the Anglo group. According to the criterion established for this analysis, these items would be categorized as biased. None of the items was identified as biased against Blacks, and only one item was categorized as biased against Chicanos (Item 22 on Vocabulary). On the other hand, nearly one-third of the items included in this analysis would be considered biased against Papagos. All of the items identified by this analysis were items on Verbal subtests, and the majority were items in the latter portions of the subtests. Three items, however, were among the earlier items of their respective subtests (Items 7 and 13 on Information and Item 6 on Arithmetic).

Point Biserial Correlations

Point biserial correlations were computed for the subtests of Information (Table 7), Arithmetic (Table 8), and Picture Completion (Table 9), as these are the only subtests on which all items are scored 1 or 0.

The correlations followed a similar pattern for all four groups on each of the three subtests. The correlations which were quite low for the beginning items, increased steadily to a peak on items in the middle portions of the subtests before dropping to fairly low points again on the upper extreme items. The low correlations on the extreme items indicate that the differences in average raw scores on the subtests were small between those who passed and those who failed these items. The items which were eliminated clearly had extremely low correlations with total subtest scores as would be expected since these items were passed or failed by 95% of the participants. Nevertheless, at least half of the items on the three subtests were discriminating well for the four groups with point biserial correlations above .40.

The correlations were lower for Papagos on Items 13, 15, 16, 17, 18, 19, 20, 21, 22, and 24 of Information and on Items 13, 15, and 17 of Arithmetic which were identified as outliers by the previous analysis. The lower point biserials indicate that these items were becoming more difficult for Papagos. Since these items were beyond the peaks on their respective subtests, this would suggest that the Papagos were reaching the ceiling on these items. At the same time, Items 13, 15, 16, 17, 18, 19, and 20 on Information had correlations with subtest scores above .40, so these items were still functioning well.

Item 7 on Information and Item 6 on Arithmetic identified as biased by the previous analysis had relatively high point biserial correlations. According to the outlier analysis, they demonstrated an item by group interaction in that they were relatively more difficult for Papagos than for Anglos. At the same time, these items discriminated quite well between those obtaining higher and lower scores. The point biserial correlations therefore tend to corroborate the characterization of these items as biased.

Table 7

Point Biserial Correlations for Information

Item	Anglos	Blacks	Chicanos	Papagos
		Information		
1	0.000	0.000	0.180	0.081
2	0.108	0.020	0.182	0.102
3	0.000	0.247	0.161	0.242
4	0.365	0.462	0.490	0.535
5	0.362	0.549	0.556	0.582
6	0.279	0.613	0.487	0.339
7	0.471	0.715	0.644	0.750
8	0.461	0.615	0.532	0.526
9	0.553	0.662	0.633	0.688
10	0.571	0.721	0.687	0.693
11	0.646	0.736	0.659	0.655
12	0.703	0.758	0.727	0.698
13	0.645	0.691	0.646	0.579
14	0.644	0.524	0.621	0.577
15	0.599	0.621	0.594	0.454
16	0.691	0.499	0.577	0.477
17	0.725	0.480	0.496	0.536
18	0.521	0.422	0.430	0.456
19	0.815	0.583	0.681	0.596
20	0.648	0.514	0.518	0.409
21	0.737	0.483	0.615	0.388
22	0.522	0.335	0.456	0.324
23	0.555	0.380	0.431	0.384
24	0.541	0.343	0.336	0.243
25	0.509	0.319	0.401	0.208
26	0.616	0.257	0.344	0.000
27	0.392	0.333	0.346	0.241
28	0.533	0.252	0.439	0.208
29	0.249	0.228	0.181	0.000
30	0.284	0.145	0.000	0.000

Table 8

Point Biserial Correlations for Arithmetic

Item	Anglos	Blacks	Chicanos	Papagos
		Arithmetic		
1	0.138	0.196	0.000	0.223
2	0.097	0.312	0.135	0.413
3	0.280	0.326	0.250	0.513
4	0.604	0.688	0.622	0.756
5	0.248	0.356	0.236	0.437
6	0.292	0.516	0.346	0.683
7	0.252	0.565	0.539	0.609
8	0.581	0.738	0.629	0.732
9	0.662	0.712	0.702	0.739
10	0.722	0.741	0.687	0.625
11	0.735	0.733	0.724	0.633
12	0.701	0.698	0.703	0.572
13	0.674	0.715	0.689	0.440
14	0.701	0.598	0.637	0.477
15	0.557	0.385	0.544	0.264
16	0.633	0.465	0.397	0.335
17	0.502	0.334	0.325	0.117
18	0.370	0.258	0.260	0.000

Table 9

Point Biserial Correlations for Picture Completion

Item	Anglos	Blacks	Chicanos	Papagos
	Picture	Completion		
1	0.000	0.000	0.000	0.187
2	0.087	0.284	0.209	0.201
3	0.000	0.350	0.140	0.153
4	0.074	0.299	0.137	0.353
5	0.276	0.401	0.350	0.258
6	0.313	0.506	0.365	0.314
7	0.404	0.482	0.439	0.397
8	0.497	0.479	0.387	0.487
9	0.549	0.595	0.538	0.493
10	0.550	0.615	0.564	0.604
11	0.643	0.732	0.568	0.478
12	0.598	0.675	0.464	0.580
13	0.650	0.603	0.625	0.558
14	0.614	0.622	0.486	0.557
15	0.607	0.688	0.603	0.549
16	0.690	0.691	0.627	0.593
17	0.685	0.686	0.681	0.609
18	0.657	0.663	0.695	0.622
19	0.616	0.573	0.588	0.642
20	0.724	0.688	0.682	0.633
21	0.521	0.344	0.448	0.435
22	0.584	0.415	0.492	0.385
23	0.608	0.453	0.482	0.469
24	0.483	0.362	0.348	0.287
25	0.481	0.321	0.408	0.355
26	0.302	0.156	0.229	0.199

DISCUSSION

Cultural bias against ethnic minority persons in individual intelligence tests has been attributed to the interpretation of test results, to factors in the testing situation, and to the individual items which comprise the test. Few studies, however, have attempted to examine evidence for item bias empirically.

Only two studies, one by Sandoval (1979) and the present one, have investigated item bias on the WISC-R among children from different ethnic and cultural groups. The results of the present investigation for Anglo, Black, and Chicano children essentially replicate the findings of Sandoval for these three groups. Since these were the only groups in the Sandoval study, the results for Native American Papago children in the present study cannot be compared. The lower mean performance of Black and Chicano children, the high and comparable internal consistency reliabilities for the three groups, and the high rank order correlations of item difficulties are very similar results in the two studies. The rank order correlations of differences in the difficulty of adjacent items are slightly lower than those presented by Sandoval. The discrepancy may be due to the use of different measures of item difficulty. Sandoval used item means as a measure of item difficulty whereas p-values were used in the present study.

The items identified by Sandoval's multivariate analysis as those most contributing to the observed differences in ethnic group performance were, with few exceptions, items in the middle portions of the subtests. The higher point biserial correlations among the middle items on the subtests of Information, Arithmetic, and Picture Completion in the current study are compatible with his analysis.

The preceding analyses are associated with the approach to the investigation of item bias advocated by Arthur Jensen (1978). On the basis of these analyses, one would conclude that these overall psychometric characteristics of the test were similar for the Anglo, Black, and Chicano groups in both studies and thus that the test was not biased.

The present investigation is the only one to date that has also employed transformed item difficulties and a method of outlier analysis for the identification of specific items as biased. The definition of bias underlying these methods is one of an item by group interaction. Items which are differentially more difficult for one group than another are considered biased.

The transformed item difficulties indicate that the majority of items on the subtests analyzed in the present study are somewhat more difficult for Blacks, Chicanos, and Native American Papagos than for Anglos. According to the criterion established in the outlier analysis, however, none of the items was biased against Blacks. Only one item was so identified for Chicanos. However, nearly one-third of the items included in the analysis would be considered biased against Native American Papagos.

Therefore, according to these results, there is no support for the contention that items on these subtests are biased against Black and Chicano children. In particular, those items frequently selected for criticism, such as Item 6 on the Comprehension and "Who invented the electric light bulb?" on Information, are not markedly more difficult for Blacks and Chicanos. The latter item is noticeably more difficult for Papagos, but the item "Who discovered America?" is not, contrary to the prediction of the critics. However, there is slight support for the notion that the Performance subtests may be less biased than the Verbal subtests since no item was identified as biased against any group on Picture Completion.

At the same time, there is evidence of bias among the items of the Verbal subtests against Native American Papago children. The biased items are, with few exceptions, in sequence in the latter portions of the subtests. Thus, as the items are becoming more difficult for all groups, they are becoming even more so for Papagos. However, these items are not substantially more difficult as can be seen in the scatterplots for Information for example. They are only slightly outside the fixed limit of difficulty used in the outlier analysis.

The important question which remains is whether the results obtained by employing these methods of identifying biased items constitute evidence of cultural bias against Native American Papago children. The greater difficulty of these items would have to be due solely to cultural/ethnic differences for this to be the case. Since cultural differences are associated with differences in socioeconomic status in American society, any statement regarding item bias due to cultural differences would have to include an analysis of the effect of socioeconomic variables. Unfortunately, the sample sizes were not large enough in the present study to permit the adequate representation of different socioeconomic levels for this analysis. The lower socioeconomic status of ethnic minorities makes sufficient representation at the upper levels difficult to obtain and will limit the possibility of adequate research on this question.

Another limitation of the present study is that no analysis was made of item difficulty while attempting to equate the groups for levels of total score. Once again, adequate representation of all groups at the upper score levels is a problem. There is also the conceptual difficulty of equating groups on total scores which consist of items suspected of bias.

Finally, there is the issue of whether items identified as biased by these methods meet an implicit assumption of those alleging cultural bias in specific items. This assumption may be stated as one of interference. The teachings of the minority culture are assumed to conflict with the content of items to make the learning of the "correct" majority culture response more difficult. The results for Papagos, as well as for Blacks and Chicanos, appear to indicate an overall pattern of greater difficulty. One might speculate, therefore, that these results point to a general lack of familiarity with the contents of these items. However, a majority of the items identified as biased against Papagos cannot be related to specific aspects of Papago culture which would interfere with the correct response. One of the biased items on information is a possible exception. Item 7 on Information involves conceptions of time which may differ in the two cultures (Joseph, Spicer, & Chesky, 1949).

On the basis of his analyses, Sandoval (1979) concluded that the results did not support the allegation of item bias on the WISC-R for Blacks and Chicanos. Sandoval based his argument on the notion that for bias to be present, there would have to be "a number of items with radically different difficulties for children from different ethnic groups" (Sandoval, 1979, p. 925). He did not specify the number nor define how "radically different" the difficulties should be. The transformed item difficulty approach with the accompanying method of outlier analysis used in the present study was intended at least to provide a means of establishing limits to define how different the item difficulties needed to be in order to characterize items as biased.

The results of both studies reveal that a majority of items spread throughout the entire test are more difficult for children of ethnic minority groups. However, as Sandoval pointed out, and as the method of outlier analysis operationalized, the existence of differences in item difficulty, _per se_, does not constitute evidence of bias. Furthermore, it is also suggested here that the presence of item by group interactions on the Verbal subtests for Papagos, although necessary, does not provide sufficient evidence of cultural bias in items in view of the possible confounding effects of socioeconomic status and the lack of any demonstrable cultural interference on a substantial number of specific items.

These results, similar to other studies of bias in tests, suggest negligible or no bias against Blacks and Chicanos. The broader concern among minority critics, is not just test bias, but rather, how test _use_ affects the lives of children (Reschly, 1981). Tests which are unbiased according to technical criteria _may_ be useful in educational classification and programming. The critical issue, however, is the quality of the special programs and interventions which sometimes follow test use.

REFERENCES

Anastasi, A. *Psychological testing* (4th ed.). New York: Macmillan, 1976.

Angoff, W. H., & Ford, S. F. Item-race interaction on a test of scholastic aptitude. *Journal of Educational Measurement*, 1973, *10*, 95-105.

APA Monitor. Exhibit A: IQ Trial. Plaintiffs Take the Stand, December, 1977.

Cleary, T. A., & Hilton, T. L. An investigation into item bias. *Educational and Psychological Measurement*, 1968, *8*, 61-75.

Diana v. California State Board of Education. No. C-70 37 RPF, District Court of Northern California (February 1970).

Flaugher, R. L. The many definitions of test bias. *American Psychologist*, 1978, *33*, 671-690.

Gay, G., & Abrahams, R. D. Does the pot melt, boil, or brew? Black children and white assessment procedures. *Journal of School Psychology*, 1973, *11*, 330-341.

Guadalupe v. Tempe School District (F. August 1971, U. S. District Court of Arizona).

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

Jensen, A. R. The current status of the IQ controversy. *Australian Psychologist*, 1978, *13*, 7-27.

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

Joseph, A., Spicer, R. B., & Chesky, J. *The desert people*. Chicago: The University of Chicago Press, 1949.

Laosa, L. M. Historical antecedents and current issues in nondiscriminatory assessment of children's abilities. In Thomas Oakland (Ed.). *Psychological and educational assessment of minority children*. New York: Bruner/Mazel, 1977.

Larry P. v. Riles. 343 F. Supp. 1306 (ND Cal 1972).

Larry P. v. Wilson Riles, C-71 2270 RFP, United States District Court, Northern District of California, 1976.

Larry P. v. Wilson Riles. Opinion U. S. District Court for Northern District of California (No. C-712270 RFP), October 11, 1979.

Mercer, J. R., & Brown, W. C. Racial differences in IQ: Fact or artifact. In Carl Senna (Ed.). *The fallacy of IQ*. New York: The Third Press, 1973.

Mercer, J. *System of Multicultural Pluralistic Assessment*. New York: Psychological Corporation, 1979.

Merz, W. R., & Rudner, L. M. Bias in testing: A presentation of selected methods. Paper presented at the meeting of the American Educational Research Association, Toronto, March, 1978.

PASE v. Hannon, U. S. District Court, Northern District of Illinois, Eastern Division, No. 74 (3586), July, 1980.

Reschly, D. Psychological testing in educational classification and placement. *American Psychologist*, 1981, *36*, 1094-1102.

Reschly, D. J., & Jipson, F. Ethnicity, geographic locale, age, sex, and urban-rural residence as variables in the prevalence of mild retardation. *American Journal of Mental Deficiency*, 1976, *81*, 154-161.

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

Sattler, J. *Assessment of children's intelligence and special abilities.* (2nd ed.). Boston: Allyn & Bacon, 1982.

Scheuneman, J. A method of assessing bias in test items. *Journal of Educational Measurement*, 1979, *16*, 143-152.

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

Williams, R. Danger: Testing and dehumanizing black children. *The School Psychologist*, 1971, *25*, 11-13.

Williams, R. The BITCH-100: A culture specific test. *Journal of Afro-American Issues*, 1975, *3*, 103-116.