An Examination of Culture Bias in the Wonderlic Personnel Test*

ARTHUR R. JENSEN

University of California, Berkeley

Internal evidence of cultural bias, in terms of various types of item analysis, was sought in the Wonderlic Personnel Test results in large, representative samples of Whites and Blacks totaling some 1,500 subjects. Essentially, the lack of any appreciable Race × Items interaction and the high interracial similarity in rank order of item difficulties lead to the conclusion that the Wonderlic shows very little evidence of cultural bias with respect to the present samples which, however, differ appreciably in mean scores. The items which account for the most variance *within* each racial group are, by and large, the same items that show the largest interracial discrimination.

Psychometricians are generally agreed that a population difference in average test score is not, by itself, evidence of biased sampling of test items such as to favor (or disfavor) a particular cultural group. The mean difference between groups may be explainable in terms of factors other than culture bias in the item content of the test. Evidence of culture bias thus depends upon criteria other than a group mean difference.

There are two main classes of criteria for assessing test bias: external and internal. They are complementary. The external criteria are the more important in terms of the practical usefulness of the test and where predictive validity for a specific quantifiable performance criterion is possible. Bias is indicated when two (or more) populations show significantly different regressions of criterion measures on test scores. If the regression lines for the two (or more) groups do not differ significantly in intercept and slope, the test can be said to be "fair" to all groups with respect to the given criterion of external validity. Refinements and variations of this general external criterion for assessing test bias have been discussed extensively in the measurement literature (e.g., Cleary, 1968; Darlington, 1971; Humphreys, 1973; Jensen, 1968; Linn, 1973; Thorndike, 1971).

Internal criteria of cultural bias become important when discussing the construct validity of the test and in assessing claims of bias even when the external validity criteria give no evidence of bias. Such claims of test bias are sometimes made on the grounds that the external criterion of the test's validity is itself culture-biased and is therefore predictable by a culture-biased test. Internal

*Requests for reprints should be sent to Arthur R. Jensen, Institute of Human Learning, University of California, Berkeley, CA 94720.

criteria of bias get around this argument by examining the degree to which different socioeconomic and cultural groups differ in terms of various ''internal'' features of the test involving item statistics. The main criterion for the detection of bias lies in the magnitude of the Groups × Items interaction relative to other sources of variance in an analysis of variance (ANOVA) design comprised of Groups (G), Items (I), Subjects within Groups (S), and the interactions G × I and S × I. This method was first used by Cleary and Hilton (1968), who examined the G × I interaction on two forms of the Preliminary Scholastic Aptitude Test in White and Black groups. The Race × Items interaction proved statistically significant but contributed so minimally relative to the main effects that the authors concluded: ''. . . given the stated definition of bias, the PSAT for practical purposes is not biased for the groups studied'' (p. 74).

The Groups × Items interaction is analyzable into two effects: (a) the similarity in the rank order of the percentage passing, p, each item in each of the groups, and (b) the similarity between the groups in the differences between the p values of adjacent items in the test, that is, $p_1 - p_2$, $p_2 - p_3$, etc. These are here called p decrements. Group differences in rank order of item difficulties are termed disordinal interactions. Group differences in p decrements, when the rank order of p values is the same in both groups, are termed ordinal interactions. A measure of similarity between groups, such as the Pearson correlation between the groups, in p values and p decrements, can serve as sensitive indexes of the degree to which the groups behave differently with respect to different items. All of the items in any omnibus test are not likely to be equally culture biased, and to the degree that items differ in this property, the extent of cultural differences between two groups relevant to performance on the test should be related inversely to the size of the intergroup correlations of p values and of p decrements. Also, if more test items are culturally relevant or unreliable in one group than in another, this can be expected to result in different magnitudes of the test's internal consistency reliability in the two groups.

The present study examines the Wonderlic Personnel Test (WPT) for evidence of culture bias in terms of these internal criteria when applied to representative White and Black samples. The WPT is an obviously culture-loaded test of general intelligence. The fact that it is culture-loaded only means that most of the items are based on specific information and cognitive skills that are commonly acquired in present-day English-speaking western culture. This is obvious simply from inspection of the test items. Whether the obvious culture loading of the items biases the test to the disadvantage of any particular population with respect to another population is a separate question that can be answered only in terms of empirical investigation of test data from the groups in question.

The cultural-educational loading of the Wonderlic would seem to make it suspect as a possibly culture-biased test in the American Black population. This should be a point of concern when the WPT is used in business and industry, and especially where precise external criteria of the WPT's validity in the White and

Black groups are not available. More than 6,500 organizations routinely use the WPT as a part of their personnel selection and placement procedures, making it one of the most widely used tests of mental ability.

Detailed descriptions of the WPT and references to previous research can be found in Buros (1972, pp. 724–6). Briefly, the WPT is a group-administered paper-and-pencil test of 50 verbal, numerical, and spatial items arranged in spiral omnibus fashion. It is generally given with a 12-minute time limit. Alternate form reliabilities average .95. Use of the WPT is claimed to have validity where educability or trainability is a job requirement (Wonderlic & Wonderlic, 1972, p. 60). Large representative samples of males and females show no significant difference in total raw score on the WPT.

Black Norms

Norms based on 38,452 Black job applicants have been published (Wonderlic & Wonderlic, 1972). The authors state: "The vast amount of data studied in this report confirms that a very stable differential in raw scores achieved by Negro applicant populations exists. Where education, sex, age, region of country and/or position applied for are held constant, Negro-Caucasian WPT score differentials are consistently observed. These mean score differentials are . . . about one standard deviation apart when comparisons of Caucasians and Negroes are studied" (p. 3). As the authors note (p. 68), the Black (as well as White) norms are based on biased samples of the Black (and White) populations to the extent that they are based on an applicant population of individuals who are looking for jobs. The age group from 20 to 24 is predominantly represented for both sexes and for both races.

The published norms show the mean and median test score of Black and White applicants for each of 80 different occupational categories, from the professional-managerial level to unskilled labor. The correlation between the Black and White medians across the 80 occupational categories is .84 (the correlation between means is .87), indicating a high degree of similarity between the racial groups in their self-selection for various occupations. In other words, the rank order of median and mean test scores of applicants for various jobs is very similar in the Black and White populations, despite the approximately 1 σ race difference in mean scores for all job categories.

Is there internal evidence in the test data that the 1 σ difference between Whites and Blacks is attributable in whole or in part to culture bias in the WPT?

METHOD

Subjects

Parallel analyses were performed on two pairs of White and Black samples. Thus the findings from the main analyses are replicated in two sets of White–Black comparisons based on samples selected in different ways.

Sample 1 consists of 544 White and 544 Black subjects representing a random sample of the nationwide population of job applicants on which the published White and Black norms are based for Form IV of the WPT, which have been given full statistical description in the manual of norms of the WPT (Wonderlic & Wonderlic, 1972). For unknown reasons, this Black sample is 4.75 raw score points below the mean of national norms for Blacks. The samples were drawn without selection for characteristics such as age, education, job category, sex, and region. All subjects coded as "other minority" or subjects with Spanish surnames were excluded from the sample. In terms of the White σ (standard deviation), the mean scores of the White and Black samples differ by 1.05 σ as compared with 1.00 σ in the total normative populations.

Sample 2 consists of randomly selected test protocols of 204 White and 204 Black subjects who were job applicants for entry level positions in a single company in New York City. No selection was made on age, education, and sex. Subjects coded as "other minorities" and Spanish surnames are not included in the White sample. The White and Black means of Sample 2 are very close to the national norms. (*Sample 2:* White $\bar{X}$ = 22.07, SD = 7.43; Black $\bar{X}$ = 15.63, SD = 6.95. National Norms: White $\bar{X}$ = 23.32, SD = 7.50; Black $\bar{X}$ = 18.50, SD = 7.06.) Kuder-Richardson (Formula 21) reliabilities are .88 in the White and .86 in the Black samples.

RESULTS

p Values and p Decrements

The *p value* is the proportion of the total sample who answer a given test item correctly; *p* values were obtained for items 1–50 in the White and Black groups.

The *p decrement* is the difference between the *p* values of ordinally adjacent test items, for example, $p_1 - p_2$, $p_2 - p_3$, etc., where the subscript indicates the item number in the test. *p* decrements between adjacent items 1–2, 2–3, . . . , 49–50 were obtained in both samples.

Table 1 shows the mean *p* values within sets of 10 items (and for all items) for each of the racial groups in Samples 1 and 2. The item *p* values were correlated between racial groups within 10-item sets and over all 50 items. (Since item difficulties gradually increase from the first to the last items of the test, correlations between item *p* values of Whites and Blacks based on the entire range over 50 items would be a much less sensitive index of the groups' similarity than correlations based on subsets of 10 items, in each of which the *p* values differ relatively little, compared to the total range of *p* values, within each racial group.) As can be seen in Table 1, these correlations are quite high even within sets of 10 items. This means that the relative difficulty of the items, as indicated by the proportion passing, is highly similar to the White and Black samples.

The size of the between-race correlations of *p* values can be viewed in relation

TABLE 1
Summary of Wonderlic Item Statistics on Sample 1 (N = 544 Whites and 544 Blacks) and Sample 2 (N = 204 Whites and 204 Blacks)

Items	Mean p				Correlation between White p × Black p		Correlation between W × B p decrements	
	Sample 1		Sample 2					
	White	Black	White	Black	Sample 1	Sample 2	Sample 1	Sample 2
1–10	.815	.623	.829	.653	.886	.920	.907	.934
11–20	.662	.409	.682	.485	.802	.702	.845	.768
21–30	.461	.233	.439	.266	.879	.945	.855	.928
31–40	.238	.101	.230	.143	.937	.943	.673	.789
41–50	.035	.007	.031	.014	.765	.933	.765	.938
All Items	.442	.275	.442	.312	.932	.956	.792	.832

to the within-race split-half correlations, which were obtained by intercorrelating the p values based on randomly split halves of each racial group. The between-race correlations (Table 1: White p × Black p) are slightly lower than the within-race correlations, which do not differ significantly for Whites and Blacks and average .90, ranging from .83 to .96 and paralleling the between-race correlations. The within-race correlation over all items is .98 for both Whites and Blacks in both Samples 1 and 2, as compared with the between-race correlations of .93 and .96.

The p decrements were treated the same way. Since p decrements, unlike p values, are not systematically correlated with the item's ordinal position in the test, the interracial correlation between p decrements is a much more sensitive index of group similarity than the correlation of p values. For example, in another study (Jensen, 1974) it was found that although the p values of two parallel forms of a test (Peabody Picture Vocabulary) showed a near perfect correlation, the correlation between the p decrements was not significantly greater than zero. A high interracial correlation between p decrements means that the relative *differences* in difficulty between adjacent items are much alike in the two racial groups. If some items were more racially-culturally biased than others, resulting in different relative difficulties for Whites and Blacks, it would be reflected in a low interracial correlation between p decrements. As can be seen in Table 1, this is not the case. Though quite high, the interracial correlations of p decrements are predictably lower than the correlations of the p values themselves, since differences between measures are always less reliable than the measures themselves.

Yet the within-race split-half correlations are remarkably high for p decrements and are practically identical for the two races, averaging .89 and ranging from .75 to .96. The within-race correlation of p decrements over all items is the same in both races: .92 in Sample 1 and .94 in Sample 2; these correlations within races may be compared with the corresponding correlations of .79 and .83 between races.

White–Black Differences According to Type of Items

It is often claimed that Blacks perform relatively less well on verbal items than on other types, since presumably verbal content allows wider scope for cultural variations and the effects of bias on Black scores. To see if this notion holds true for the various kinds of item content in the WPT, items were classified by the writer as verbal (V), numerical reasoning (N), and logical reasoning (R). (*Verbal:* Items 1, 2, 3, 4, 6, 7, 9, 10, 11, 13, 15, 16, 17, 18, 19, 20, 21, 23, 25, 26, 28, 30, 32, 33, 37, 39, 42, 46; *Numerical:* Items 5, 8, 22, 29, 34, 36, 38, 41, 43, 45, 47, 49, 50; *Reasoning:* Items 12, 14, 24, 27, 35, 40, 44, 48.) There are 27 V, 13 N, and 8 R items. (Two items, classified as "factual information" and "clerical accuracy," were not used in this analysis.) Every subject's test was scored so as to yield V, N, and R scores, expressed as the percentage correct within each scale. The White–Black differences on each of the scales, expressed in σ units (i.e., the mean difference/average standard deviation within groups) are $V = 0.93$, $N = 0.78$, $R = 1.03$. An analysis of variance shows that the Race × Scale interaction is significant, F (2,812) = 15.43, $p < .01$. A Scheffé test of all the possible contrasts shows that the significant interaction is due to the N scale; Whites and Blacks differ significantly less on the numerical items than on the verbal and logical reasoning items, which do not differ significantly in degree of racial discrimination. Though the Race × Scale interaction is significant, it should be noted thPt its mean square variance (1759.5) is less than one-twentieth of the mean square variance of the main effect of race (40042.3).

Is the interaction of VNR scales with race a truly racial (or cultural) effect or is it due to an interaction between VNR and overall ability level? To answer this question, White and Black groups were selected so as to have identical means and standard deviations on the total Wonderlic score. Perfect matching of total scores was possible on 127 White–Black pairs. The same analysis of variance as was performed on the total samples (above), was performed on the two racial groups equated for total scores. It shows a nonsignificant interaction of Race × VNR scales, F (2, 504) = 1.57. That is to say, Whites and Blacks equated for total score do not differ significantly on the Verbal, Numerical, and Reasoning scales. This finding suggests that the significant Race × VNR scales interaction found in the total sample is really not a Race × VNR scales interaction but may be due to the interaction of General Ability Level × VNR scales. To test this hypothesis, the total White sample was divided into two groups (each with N = 98), one with approximately the same total Wonderlic score distribution as the White norms and the other with approximately the same score distribution as the

Black norms. These two groups of Whites, duplicating the White and Black population differences, are here referred to as "pseudo-racial" groups. The same analysis of variance as in the two previous analyses was performed on the two pseudo-racial groups to test the significance of the Pseudo-race × VNR scales interaction. The interaction is nonsignificant ($F < 1$). This can only mean that the significant Race × Scales interaction found in the total samples of Whites and Blacks is neither purely a Race × Scales interaction nor a General Ability × Scales interaction, but involves a more complex triple interaction of Race × General Ability Level × Scales.

To test this triple interaction each of the total racial groups was divided at its own median on the Wonderlic total score into higher and lower ability levels. A three-way ANOVA was performed, with the main effects of Race, Ability Levels, and VNR Scales and all of their possible interactions, and with subjects nested in Race and Ability Levels. The Race × Ability Levels × VNR Scales interaction is significant (F (2, 808) = 5.34, $p < .01$). The triple interaction results from the fact that within the White sample the higher and lower ability levels show no significant interaction with VNR scales, while in the Black sample the higher and lower ability levels differ significantly more on the Verbal scale than on the Numerical and Reasoning scales. Although this triple interaction is statistically significant, it is trivial in terms of the proportion of variance accounted for, amounting to less than 0.2% of the total variance. The race main effect has a mean square variance (with 1 *df*) 17 times greater than the total of the mean square variances of the first- and second-order interactions (with 4 *df*) involving Race and VNR scales (i.e., Race × Scales, and Race × Ability Levels × Scales).

Can psychologically sophisticated judges identify test items that discriminate most between Whites and Blacks? To find out, an index of intergroup discriminability was obtained for every one of the 50 test items. The index used here, the Z index of item discriminability, is explained by Guilford (1954, pp. 418–419). It expresses the proportion p of a group passing an item in terms of the Z score deviation of the normal curve. The intergroup difference for a given item, then, is expressed as the difference between their Z scores. By thus transforming p values to Z scores, items of different absolute difficulty in the two racial groups can be ordered on an interval scale. This was done, and the 8 most and the 8 least racially discriminating items were selected. (Most discriminating items, Nos. 5, 6, 7, 8, 11, 16, 18, 49; least discriminating, Nos. 2, 20, 26, 27, 34, 36, 44, 48.) The items were typed, without any identification, on separate cards, and 10 judges were asked to sort the cards into two piles of 8 cards each. The judges were instructed that half of the 16 items were found to be the most racially discriminating and half were the least discriminating in a test of 50 items. Two of the judges had PhDs in psychology and the others were advanced graduate students working for PhDs in psychology. There were five

Whites and five Blacks. The percentage of correct classifications for each judge was determined. Only one judge (Black) did better than the chance score of 50%, with 62.5% correct classifications. The overall mean of the 10 judges was 38.75% correct. White judges averaged 32.5%, Blacks 45%. Thus both groups did somewhat worse than chance. Apparently the most and least racially discriminating items are not at all easily identifiable by the subjective judgments of either Whites or Blacks with a background in psychology.

It was hypothesized that items' loadings on the first principal component when the item intercorrelation matrix is factor analyzed (i.e., a principal components analysis with unities in the principal diagonal) within each racial group separately would be most highly related to the item's discriminability between the racial groups. That is to say, the more highly an item is correlated with the most general factor common to all items, *within either* racial group, the more highly it will discriminate *between* the racial groups. This prediction is, of course, the opposite of what one would predict from a culture bias hypothesis, according to which the items that discriminate the most between the races should not be the items that account for most of the variance *within* races. (The extreme case of culture bias would be the test item to which the correct answer is known by all the members of one cultural group and by no members of some other cultural group.)

To test this hypothesis, the items' loadings on the first principal component of the item intercorrelation matrix were obtained from separate principal components analyses of the White and Black data. The first principal component is a factor which accounts for more of the variance than any other factor and can be regarded as the general factor of the Wonderlic test items. The items' loadings on the first principal component were correlated with the items' Z index of interracial discriminability for all items. The Pearson correlation is .47 in the White sample and .62 in the Black sample. These correlations should be evaluated in terms of the "reliability" of the pattern of the factor loadings in the White and Black samples, and the "reliability" of the pattern of the items's index of interracial discriminability. To determine this, the two racial samples were each randomly split in half and the first principal component was extracted for each half-sample. The correlation of the 49 factor loadings across the split halves of the White sample is .69, and of the Black sample, .73. The correlation of the factor loadings across Whites and Blacks (total samples) is .68. Thus it is apparent that the correlation of factor loadings between races is about the same as the correlation within races. The items's Z index of racial discriminability shows a split-half samples correlation of .49. In other words, the within-race reliability of the pattern of the item loadings on the first principal component (or general factor) is about .70 and the within-race reliability of the A index of racial discriminability is .49. (The reliability of the discriminability index is much lower probably because it is based on the profile of group differences over the 49 test items, as well as the fact that many of these differences are highly similar.) Given these reliabilities, the correlations of .47 (for Whites) and .62 (for Blacks) between items' g loadings and their racial discriminability must be regarded as substantial, since the theoretically highest correlation one could expect to find would be about $(.70 \times .49)^{1/2} = .59$. (I have purposely avoided stepping-up the split-half "reliabilities" by means of the Spearman–Brown formula in this case, because the statistical assumptions underlying this procedure are extremely problematic in the case of the profile of factor loadings. Treatment of this problem is beyond the scope of the present paper. The split-half reliability of the index of racial discriminability, however, could be legitimately stepped up by $2r/(1+r)$ without violating the statistical assumptions in the Spearman-Brown formula.)

In short, there is a substantial relationship between the size of the item loadings on the general factor common to all items in tiie Wonderlic and the magnitude of the White–Black difference on the item, and this is true whether the general factor is determined in the White or in the Black sample. The items that best measure the general factor *within* each racial group are the same items, by and large, that discriminate most highly *between* the racial groups.

Analysis of Variance: Items × Subjects Matrix

The Race × Items interaction in a complete ANOVA of the Items × Subjects matrix provides a sensitive index of item bias relative to other sources of variance. Using the Sample 2 data, three such ANOVAs were performed: (1) on the total White and Black groups, (2) on White and Black groups equated on total WPT score, and (3) on "pseudo-racial" groups comprised entirely of two groups of White subjects selected so that their total WPT score distributions closely match the normative White and Black distributions in means and SDs. The ANOVAs for each of these conditions are summarized in Table 2. So that the three analyses can be directly compared, the sum of squares for each source in the ANOVA is converted to omega squared (ω^2) × 100, which is the percent of the total variance attributable to the given source.

The ANOVA is used here not as a test of significance but as a means of showing the percentage of variance (omega squared) attributable to the main effects and their interactions. More important than statistical significance for our purposes is the magnitude of the Race × Items interaction relative to other sources of variance. The larger it is, the more "unfair" the test as regards culture bias. The appropriate index of unfairness or bias, thus defined, is the B/A ratio, which, in terms of ω^2, is A = R/S and B = (R × I) / (I × S). In terms of F, $B/A = F_{\mathrm{R}\times\mathrm{I}}/F_{\mathrm{F}}$. The two formulas for the B/A ratio are algebraically equivalent. The B/A ratio provides a scale or index on which tests may be compared for culture bias, as here defined. If the F for the Race × Items interaction is less than 1, it is presumed that no bias at all has been demonstrated and there is no point in computing the B/A ratio, which can be assumed to be zero. The higher

TABLE 2
Omega Squared ($\omega^2 \times 100$) and F from ANOVA of Wonderlic Test in Total and Equated White and Negro Samples, and in "Pseudo-Race" Samples

Source of variance[a]	Total samples			Equated samples			"Pseudo-race" samples		
	df	$\omega^2 \times 100$	F^b	*df*	$\omega^2 \times 100$	F^b	*df*	$\omega^2 \times 100$	F^b
Race (R)	1	1.83	84.87	1	0	0	1	2.08	58.87
Items (I)	48[c]	34.22	256.54	48	37.95	172.86	48	39.36	150.57
Subjects within race (S)	406	8.75	7.75	252	6.45	5.56	194	6.87	6.51
R × I	48	1.04	7.83	48	0.29	1.26	48	0.94	3.61
S × I	19,488	54.17		12,096	55.31		9,312	50.73	

[a] The sum of squares (SS) for any given effect in the ANOVA can be obtained by multiplying the total SS by ω^2. The total SS are: Total samples = 4732.24, Equated samples = 2949.26, "Pseudo-race" samples = 2375.79. The mean square variance for any effect is obtained by dividing the SS for the effect by its degrees of freedom.

[b] The F ratios in this table accurately indicate the ratios of the main effects MSVs to the interaction MSVs, but they are not used here as significance tests. The p values of these Fs cannot be determined with exactitude because of the lack of independence of item means and variances. However, the relative magnitudes, rather, the level of statistical significance, are the most important aspects of this analysis.

[c] There are only 48 degrees of freedom for Items, since in this analysis Item 50 was omitted, as no subject of either race got it correct and it therefore contributes nothing to the total variance.

the value of the B/A ratio, the easier it would be to equalize or reverse the racial group means by item selection. Obviously a small group mean difference along with a large Groups × Items interaction would mean that a somewhat different selection of items from the same item population could equalize or reverse the group means. The lower the value of B/A, the less is the possibility of equalizing the group means through item selection from a similar population of items. This would not rule out the possibility of introducing different kinds of items into the test, but if doing so increases the B/A ratio (even though it decreases the group mean difference) it can be argued that the minimizing of the group mean difference is simply a result of balancing item biases. Some tests equate male and female scores on this basis, balancing items that favor one sex with the selection of items that favor the other. Such a test, resulting in little or no mean sex difference but a large Sex × Items interaction, of course, precludes the use of such a test for studying the question of sex differences in the ability which the test purports to measure. The same thing would be true of any test which was made to equalize racial group differences at the expense of greatly increasing the Race × Items interaction. The desirable condition is to minimize the interaction as much as possible.

The *B*/*A* ratio for the total samples (Table 2) is .09. For comparison, a similar study of White and Black elementary pupils showed a *B*/*A* ratio of .14 on the culture-loaded Peabody Picture Vocabulary Test and of .06 on the culture-reduced Raven's Progressive Matrices (Jensen, 1974). Thus the Wonderlic appears to be more or less intermediate between the Peabody and the Raven on this index of bias.

ANOVA on Equated White and Black Samples. In a previous study, it was found that when groups of White and Black school children were roughly matched for mental age (rather than chronological age), and ANOVA of the Peabody Picture Vocabulary Test (PPVT) items was performed, the Race × Items interaction was greatly reduced from its magnitude when the two racial groups were of the same chronological age but different mental ages (Jensen, 1974). This finding suggests that a large part of the Race × Items interaction is attributable to a Mental Maturity × Items interaction rather than to a racial–cultural difference per se. And this hypothesis was strengthened by showing that the same magnitude of the actual Race × Items interaction could be achieved entirely with the White sample, simply by dividing it into two "pseudo-racial" groups for the ANOVA. One group of White subjects was selected so that their distribution of total PPVT scores matched the Black distribution in mean and SD; the other group of White subjects was selected so that its PPVT score distribution matched the total White distribution. When these two culturally homogeneous groups, corresponding to the Black and White samples, were subjected to the same ANOVA as was applied to the true racial groups, it reproduced the same results almost perfectly, including the Race × Items interaction. In other words, an interaction of this magnitude could be attributed to an average ability difference between the groups rather than to a cultural difference.

The same kind of analysis is here applied to the Wonderlic data. Since mental age is not a meaningful scale in an adult population, Black and White subjects were simply matched for total score on the WPT. Perfect matching was possible of 127 White–Black pairs, making the White and Black total score distributions identical.

If the WPT items are culture-biased for Blacks, one might expect that Whites and Blacks with the same total scores would obtain them in different ways, so that even when the main effect of race is zero in the ANOVA, the Race × Items interaction would remain approximately unchanged.

Table 2 shows the results of the ANOVA on the equated samples. The main effect of race was, of course, forced to be zero by equating the groups. But note that the Race × Items interaction is now very small. It is quite irrelevant to the present argument whether this interaction is or is not statistically significant. We are concerned with its magnitude relative to the interaction term in the ANOVA on the Total Samples. (A proper test of significance of the R× I interaction of the Equated Samples ANOVA would call for a different type of ANOVA [a levels design with the matched pairs as the levels], but then it would not parallel the Total Samples ANOVA.) The important point is that the R × I interaction variance is markedly reduced when the racial groups are equated on the overall test score. This finding is consistent with the hypothesis that the R × I interaction in the ANOVA of the Total Samples is more a function of the average difference in ability between the groups rather than of any cultural difference. It seems less likely that equating the White and Black groups for total score should wipe out an R × I interaction if it truly reflected a cultural difference between White and Black groups.

One might argue that White and Black subjects who attain the same total score must be highly similar in cultural background and therefore would show no significant R × I interaction. But are they culturally more similar than individuals of the same racial group who differ by 7 points in total Wonderlic score? (The σ of total scores in the normative White population is close to 7.) Siblings reared together in the same family differ by almost as much. Since the White and Black population means differ by close to 1 σ (or 7 points on the WPT), we can do an ANOVA on a "pseudo-race" comparison by making up two groups of White subjects selected so that their score distributions closely approximate those of Blacks and Whites. This was done by ranking all White scores from highest to lowest, and then, working in from both ends of the distribution, selecting pairs of subjects who differ by exactly 7 points in total score.

Table 2 shows the ANOVA of these "pseudo-race" groups. It can be seen that the results resemble the true racial comparison (Table 2—Total samples), especially as regards the R × I interaction, which for the Total samples constitutes 1.04% of the variance and for the "pseudo-racial" samples is 0.94%. The *B*/*A* ratios for the Total sample and "pseudo-race" sample are .09 and .06, respectively. The ratio of ω^2 for the interactions (R × I) / (S × I) is exactly the same (.019) in both the Total sample and the "pseudo-race" sample. All this indicates that a large part of the R × I interaction can be attributed to a Level-of-ability × Items interaction, since it is shown to exist in the two "pseudo-race" groups that are both comprised of White subjects differing in average ability. If the significant R × I interaction were explainable only in terms of cultural differences between the White and Black groups, it seems highly improbable that it could be greatly reduced simply by equating the racial groups for overall level of ability, or that the same size of interaction could be produced within a culturally homogeneous White sample divided into high and low ability groups with overlapping score distributions similar to the total White and Black distributions. In brief, from these three ANOVAs shown in Table 2, it would seem difficult to make a case that the Race × Items interaction is attributable to cultural bias. These analyses should have produced markedly different results if the popular claims of culture-biased test items were in fact valid.

The only counter hypothesis to explain these results is that the lower scoring

Whites in the pseudo-race comparison differ from the higher scoring Whites in the same way that Blacks differ from Whites, because the low-scoring Whites and the majority of Blacks presumably are both culturally disadvantaged and therefore share the same item biases. A neat way to test this hypothesis would be to make up pseudo-racial groups entirely of Whites in which each member of every pair placed in the high and low groups (differing by 1 σ) are siblings reared in the same family. Then the ''Pseudo-race'' × Items interaction could in no way be attributed to a cultural difference between the high and low ability groups.

DISCUSSION AND CONCLUSION

Several different analyses of test item characteristics have failed to reveal evidence of culture bias for large Black and White samples on the Wonderlic Personnel Test. If some items were more culture biased than others with respect to the cultural backgrounds of Blacks and Whites, one should expect (a) significantly different rank order of *p* values (percent passing) for various items in the White and Black samples; (b) significantly different intervals (i.e., *p* decrements) between the *p* values of adjacent test items in White and Black samples; (c) a substantial Race × Items interaction in the analysis of variance of the Race × Items × Subjects score matrix, even when both racial groups are equated for total score; and (d) systematic differences in the *types* of item content that discriminate most and least between the White and Black samples. None of these expectations was borne out by the present data. The small but significant Race × Items interaction could be greatly reduced by equating the White and Black groups for overall score, which would not be expected if the two groups differed culturally in reaction to the test items. Moreover, it was possible to produce a ''Pseudo-race'' × Items interaction within the culturally homogeneous White group, comparable to that found in the White versus Black comparison, simply by dividing the total White sample into two groups, one which duplicated the mean and SD of the Black norms and the other which duplicated the mean and SD of the White norms. This suggests that the Race × Items interaction is more an Ability Level × Items interaction rather than an interaction due to cultural differences. Whatever abilities or aptitudes the Wonderlic measures, they are measured by items that are internally consistent within both Black and White samples.

The only way one could view these findings as being consistent with the hypothesis that the Wonderlic is a culturally biased test for Blacks would be to claim that culture bias depresses Blacks' performance on all the test items to much the same degree. This would seem highly unlikely for cultural effects per se, especially considering the great variety of item content in the Wonderlic. Otherwise it should be possible to make up subscales consisting of items on which the Black group on the average does as well or better than the White group. This, however, is not possible with the present pool of Wonderlic items. The items that best measure the general factor common to all items *within* each racial group are also the same items that discriminate the most *between* the racial groups.

The present analyses yield no consistent or strong evidence that the Wonderlic is reacted to differently by Blacks and Whites, except in overall level of performance, in which the normative populations differ by about one standard deviation.

REFERENCES

BUROS, O. I. (Ed.) *Seventh mental measurements yearbook.* Vol. I. Highland Park, N.J.: Gryphon Press, 1972.

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

CLEARY, T. A., & HILTON, T. L. An investigation of item bias. *Educational and Psychological Measurement,* 1968, **28,** 61–75.

DARLINGTON, R. B. Another look at "cultural fairness." *Journal of Educational Measurement,* 1971, **8,** 71–82.

GUILFORD, J. P. *Psychometric methods,* 2nd ed. New York: McGraw-Hill, 1954.

HUMPHREYS, L. G. Implications of group differences for test interpretation. In Proceedings of the 1972 Invitational Conference on Testing Problems. *Assessment in a pluralistic society.* Princeton, N.J.: Educational Testing Service, 1973. Pp. 56–71.

JENSEN, A. R. Another look at culture-fair tests. In Proceedings of the 1968 Western Regional Conference on Testing Problems. *Measurement for educational planning.* Berkeley, Calif.: Educational Testing Service, Western Office, 1968. Pp. 50–104.

JENSEN, A. R. How biased are culture-loaded tests? *Genetic Psychology Monographs,* 1974, **90,** 185–244.

LINN, R. L. Fair test use in selection. *Review of Educational Research,* 1973, **43,** 139–161.

THORNDIKE, R. L. Concepts of culture-fairness. *Journal of Educational Measurement,* 1971, **8,** 63–70.

WONDERLIC, E. F., & WONDERLIC, C. F. *Wonderlic personnel test: Negro norms.* Northfield, Illinois: E. F. Wonderlic, 1972.