Person. individ. Diff. Vol. 8, No. 3, pp. 295–301, 1987
Printed in Great Britain

0191-8869/87 $3.00 + 0.00
Pergamon Journals Ltd

BLACK–WHITE BIAS IN 'CULTURAL' AND 'NONCULTURAL' TEST ITEMS

ARTHUR R. JENSEN[1] and FRANK C. J. MCGURK[2]
[1]School of Education, University of California, Berkeley, CA 94720 and
[2]Pompano Beach, Florida, U.S.A.

(*Received 19 March 1986*)

Summary—The original data of McGurk's classic study of black–white differences on cultural and noncultural tests is re-analyzed at the item level to investigate the role of possible item biases that would cause the noncultural items to be relatively more difficult than the cultural items for blacks than for whites. The evidence indicates that McGurk's results cannot be explained in terms of item biases, but appear to be the result of the noncultural items requiring more sheer reasoning ability than the cultural items, which depend more on acquired information.

INTRODUCTION

The overwhelming failure of research to demonstrate psychometric bias of any kind as contributing significantly to the average black–white difference on IQ tests (Eysenck, 1984; Jensen, 1980, 1984; Wigdor and Garner, 1982) renders the meaning of 'cultural' bias as highly dubious in this context. The notions of 'cultural bias' and of 'culturally loaded', with reference to tests or test items, are seldom clearly specified as to their objective meaning. Yet these notions seem to die hard as explanations of the average black–white difference in mental test scores (e.g. Scarr and Saltzman, 1982, pp. 864–865), despite the fact that no researcher, as yet, has systematically demonstrated a significantly *smaller* mean black–white difference on tests as a function of a reduced cultural loading, measured independently of the black–white difference itself, in tests which are equated in overall difficulty level separately within either the black or the white population.

Only one study in the now quite vast literature on test bias has met these methodological requirements, and its finding, surprisingly enough, is just the opposite of the common expectation based on the cultural hypothesis. This is the classic study, originally a Ph.D. dissertation, by McGurk (1951). Since various aspects of the study have been described in a number of articles (McGurk, 1953a, 1953b, 1967), we will here only summarize the essential gist of it.

McGurk collected a representative sample of 226 test items from various well-known group-administered IQ tests that were widely used at the time, such as the Otis Test, Thorndike CAVD, and the American Council on Education Test. A panel of 78 judges, including professors of psychology and sociology, educators, professional workers in counseling and guidance, and graduate students in these fields, were asked to classify each of the 226 test items into one of three categories: I, least cultural; II, neutral; III, most cultural. Each rater was permitted to ascribe his own meaning to the word 'cultural' in classifying the items. McGurk wanted to select the test items regarded as the most and the least 'cultural' in terms of some implicit consensus as to the meaning of this term among psychologists, sociologists, and educators. Only those items were used on which at least 50% of the judges made the same classification or on which the frequency of classification showed significantly greater than chance agreement. The main part of the study then consisted of comparing blacks and whites on the 103 items claimed as the 'most cultural' and the 81 items claimed as the 'least cultural' according to the ratings described. The 184 items were administered to 90 high school seniors. From these data, items classed as 'most cultural' were matched for difficulty (i.e. percentage passing) with items classed as 'least cultural'; there were 37 pairs of items matched ($\pm 2\%$) for difficulty.

These 37 pairs of matched items were then administered as a test to seniors in 14 high schools in Pennsylvania and New Jersey, totalling 2630 whites and 233 blacks. Because there were so many

more whites than blacks, it was possible for McGurk to obtain practically perfect matching of a white pupil with each of 213 black pupils. Each black pupil was paired with a white pupil in (1) the same curriculum, (2) the same school, and (3) enrollment in the same school district since first grade. The white–black pairs were also matched so that the white member of each pair was either equal to or lower than the black member on an 11-item index of socioeconomic background (the Sims Scale). (Exact matching on the 11 items of the SES index was achieved, on the average, in 66% of the 213 matched black–white pairs.) The matched black and white groups averaged 18.2 and 18.1 yr of age, respectively.

The results flatly contradicted the hypothesis that the white–black difference in test scores is due to the cultural loading of the items, at least as the culture loading of test items is commonly judged. On the test composed exclusively of the 37 items classified as 'most cultural', the mean white–black difference (expressed in units of the average standard deviation in the two samples) is 0.30σ, as compared with the mean difference of 0.58σ on the test composed of the 37 items classified as 'least cultural'. In a subset of 28 pairs of 'most' and 'least' cultural items that were matched for difficulty (based on the per cent passing in the combined samples), the mean black–white differences are 0.32σ and 0.56σ on the 'most' and 'least' cultural tests, respectively. Hence differences in item difficulty are not responsible for the relatively greater black deficit on the 'least cultural' items.

All of McGurk's analyses were based on the sets of 'most' and 'least' cultural items treated as whole tests. No attempt was made to analyze the data at the item level. In fact, statistical techniques for detecting item bias had not been developed at the time of McGurk's study. Because of the remarkable data and findings reported by McGurk, it now seems desirable to examine the data in light of modern concepts and techniques of item bias, to determine whether the tests composed of the 'most' and 'least' cultural items differ appreciably in item bias with respect to black–white differences. McGurk's original study had not considered the data from the standpoint of item bias. The purpose of the present study is to examine McGurk's data with some of the recently developed techniques used in the study of item bias.

METHOD

McGurk's original tests, henceforth labeled cultural (C) and noncultural (NC), of 37 items each, obtained on 213 black (B) and 213 white (W) high school seniors, were keypunched at the item level for computer analysis. All of the test items were multiple choice. Correct and incorrect responses were quantitized 1 and 0, respectively. The methods of analysis are best described in connection with their results.

RESULTS

Verification of original findings

Although the primary focus of the present study is on analyses that were not performed in McGurk's original study, McGurk's original statistics were re-calculated by computer. The results are identical to those originally reported by McGurk, within the limits of rounding errors beyond the last number of significant digits. The means and SDs (in parentheses) of the raw scores (number correct) on the NC and C tests for groups W and B are as follows:

Test	White	Black	W–B
NC	15.61 (5.50)	12.61 (4.57)	3.00
C	10.46 (5.35)	8.99 (4.60)	1.47
NC–C	5.15	3.62	1.53

The interaction of Test × Race, that is, the difference between the differences (1.53), is significant beyond the 0.01 level. This was McGurk's main finding, namely, that the mean black–white difference is significantly larger on the noncultural than on the cultural test. All of the other results reported henceforth are based on new analyses aimed mainly at the detection of item biases.

Reliability of C and NC tests

Since the C and NC tests have the same number of items (37), their internal consistency reliabilities (KR–10) are directly comparable as measures of item homogeneity. The Kuder–Richardson (Formula 20) reliability coefficients are as follows:

	C	NC
Black	+0.766	+0.690
White	+0.817	+0.787

Correlation between C and NC tests

The Pearson correlation (r) between total scores on the C and NC tests, and the r corrected for attenuation (using the KR-20 reliability coefficients), are as follows:

	r	r corrected
Black	+0.51	+0.70
White	+0.60	+0.75

The disattenuated correlations, which scarcely differ between the black and white samples, are very substantial, indicating that the C and NC tests measure very much the same ability. For comparison, the median correlation between the Wechsler Intelligence Scale for children and the Stanford–Binet in 47 studies is +0.80, which is raised to about +0.84 with correction for attenuation.

Correlation of C and NC scores with SES

Socioeconomic status (SES) was measured by the Sims Scale, a questionnaire of 11 items related to SES (Sims, 1927). (High scores indicate high SES). The mean and SD of the Sims scores are:

	Mean	SD
Black	4.43	1.97
White	4.16	1.93

The correlation of Sims scores with the C and NC tests are:

	C	NC	Combined
Black	+0.11	+0.19	+0.18
White	+0.35	+0.32	+0.37

Although the black and white groups are closely matched on SES, whites show significantly higher correlations between SES and test scores. (The black–white difference in correlations for the C and combined tests are significant beyond the 0.01 level; the NC correlations do not differ significantly at the 0.05 level.) It is a quite general finding that IQ is less highly correlated with SES in the black than in the white population. In the present data, this difference in correlations cannot be attributed to differences in restriction of range of ability. When the correlations between Sims scores and test scores are corrected for the lower SDs of the test scores in the black groups, the corrected correlations [using Formula 50 in McNemar (1949), p. 126] for the black group are C = +0.13, NC = +0.23, and Combined = +0.21, all of which are still smaller than the corresponding correlations in the white group. Correction of the correlations for attenuation also fails to wipe out the black–white difference in correlation between SES and test scores. This difference appears to be a real phenomenon. It has yet to be explained.

Table 1. ANOVA of cultural, noncultural and combined tests

Source of variance	df	Cultural (C) MS	Cultural (C) F	Noncultural (NC) MS	Noncultural (NC) F	Combined df	Combined MS	Combined F	η^2 (×100) C	η^2 (×100) NC	η^2 (×100) Comb.
Between race	1	6.06	9.04*	25.82	37.30*	1	28.45	26.68	0.19	0.69	0.41
Between items	36	18.20	131.71*	19.65	112.06*	73	20.17	127.14*	21.40	19.02	21.37
Between subjects (within race)	424	0.67	4.85*	0.69	3.95*	424	1.07	6.72*	9.28	7.89	6.56
Race × item	36	0.16	1.16†	0.42	2.40*	73	0.33	2.10*	0.18	0.40	0.35
Subjects × item (within race)	15,264	0.14		0.17		30,952	0.16		68.91	71.98	71.29

*$P < 0.001$. †Nonsignificant at 0.05 level.

Analyses of item bias

The key question to be addressed by the analysis of item bias is whether the NC test, composed of the items judged to be the 'least cultural', shows a significantly larger mean difference between blacks and whites than the C test, composed of items judged to be 'most cultural', because the NC test is in some way actually more biased than the C test, irrespective of the 78 judges' consensus concerning the items' cultural loading. Item bias is examined here by several methods. These have been explicated in detail elsewhere (Jensen, 1980, Ch. 9).

Analysis of variance. The main point of interest in the ANOVA of the groups × items × subjects data matrix is the interaction of groups and items (i.e. the Race × Item term in the present analysis) in relation to the main effect for groups (i.e. Race). The Group Difference/Interaction (or GD/I) ratio is the ratio of the F for the main effect of Race (R) to the F for the interaction of Race × Item ($R \times I$), i.e. GD/I = $F_R/F_{R\times I}$ (see Jensen, 1980, p. 561). Larger values of GD/I indicate lesser item bias relative to the size of the group difference, indicating lesser probability that the observed group difference could be attributable to item biases. Values of GD/I greater than 2 indicate that the mean difference between groups cannot be attributed to an unfavorable balance of Group × Item biases and could not be appreciably reduced by eliminating some items or adding new items selected at random from the same general population of items.

The ANOVAs of the C, NC, and combined items are shown in Table 1. Also included are the values of eta squared (η^2) × 100, which indicate the percentages of the total variance in the matrix attributable to each source. It is noteworthy that the C items show a nonsignificant Race × Item interaction. This means that the correlation between the relative difficulty of the 37 items in the black and white groups is not significantly less than 1. The GD/I ratios are 7.39 for C items, 15.54 for NC items, and 12.75 for the combined items. Thus the Race main effect, relative to the Race × Item interaction, is twice as great for the NC test as for the C test. For comparison, when the same kind of analysis was performed on white and black groups on the culture-reduced Raven Matrices and the culture-loaded Peabody Picture Vocabulary test, their GD/I ratios were 17.32 and 7.10, respectively (Jensen, 1980, p. 572). The culturally heterogeneous Wonderlic Personnel Test (in samples of 544 white and 544 black adults) showed a GD/I ratio of 10.84 (Jensen, 1977). Hence, it appears that reducing the commonly judged cultural loading of test items increases the test's mean black–white difference relative to the degree of item bias as indicated by the Race × Item interaction.

Rank order of item difficulty. If test items have the same meaning and cultural exposure for blacks and whites, then the difficulty of the items relative to one another should be the same in both groups, within the limits of sampling error. A simple test of this hypothesis is the rank–order correlation between the item difficulties (or conversely, the per cent passing each item) in the black and white groups. The rank correlation (ρ) has the one advantage that it is unaffected by the scale of measurement of item difficulty. For the 37 C items, $\rho = +0.981$, for the 37 NC items, $\rho = 0.966$, and for the 74 combined items, $\rho = +0.975$. It is especially noteworthy that combining the C and NC items does not lower the correlation between the rank order of item difficulty in the black and white groups. That is, the rank order of difficulty that the C and NC items assume with respect to each other is highly similar ($\rho = +0.975$) for blacks and whites. This fact and the high values of all these correlations suggest the absence of any appreciable cultural bias in either of the two tests. The mean rank (with the easiest item ranked 1) of item difficulty in the combined set of 74 items for each item type within each racial group is as follows:

	Black	White
C	42.7	43.6
NC	32.3	31.4

Therefore, in both groups the C items are more difficult than the NC items. Although McGurk had matched the C and NC items for difficulty level in an independent group of 90 high school students, they are not so closely matched on difficulty in the present groups. The mean per cent passing the items is as follows:

	Black	White
C	15.1	20.9
NC	30.2	39.9

The difference between C and NC items in difficulty level raises the question of whether their difference in black–white discriminability is due to their difference in level of difficulty. The most direct way to investigate this is to match C and NC items on percent passing in the *present* groups and compare the size of the black–white differences on only the matched items. The black–white differences on single items cannot properly be based on their differences in per cent passing, because per cent passing does not constitute an equal-interval scale of item difficulty. A normal-curve z transformation of the item p values (i.e. p = proportion passing) puts item difficulty on an interval scale (see Jensen, 1980, p. 439). Values of z have a mean of 0 and SD of 1. Average item difficulty ($p = 0.50$) transforms to $z = 0$, with positive values of z indicating greater difficulty. Hence the black–white difference on an item can be expressed as the difference in the z values for the given item. In the white sample, 27 pairs of C and NC items could be matched on per cent passing within $\pm 2\%$. In the black sample, 26 pairs of items were matched within $\pm 2\%$. The mean black–white difference (in z units) in item difficulty for the matched C and NC items is:

	White–black difference		
	NC	C	NC–C
Matched for whites	0.229	0.154	0.075
Matched for blacks	0.311	0.145	0.166

Hence even for C and NC items matched for difficulty within either racial group, the NC items show a larger mean black–white difference than the C items. Moreover, the size of the black–white difference is completely unrelated to the item difficulty of the matched items. When the black–white differences on C and NC items are correlated across the matched pairs, the correlations are +0.005 for the 26 item pairs matched in the black group and −0.071 for the 27 item pairs matched in the white group.

Correlations based on the delta scale of item difficulty. A Pearson correlation between item difficulties in the black and white groups should be based on item difficulties represented on an interval scale of difficulty. A conventional transformation of p values is the delta scale, in which Δ is simply a linear transformation of the normal z transformation of p; that is, $\Delta = 4z + 13$, giving Δ a mean of 13 and SD of 4 (see Jensen, 1980, pp. 439–440). The Pearson r between the black and white Δ values of the 37 C items is +0.957, of the 37 NC items, +0.946, and of the 74 items combined, +0.956. These correlations do not differ significantly from one another, but they all differ significantly from the corresponding average correlation (boosted by the Spearman–Brown formula) between randomly selected halves of each same-race group, which is about +0.98. This can be taken as the reliability of the profile of Δ values within each racially homogeneous group, and indicates that there is some small but statistically significant degree of black–white item bias in both the C and NC tests.

Identification of specific biased items. The item delta values within each test and racial group were transformed to delta prime, Δ', so as to have an overall mean of 13 and SD of 4 *within* each of the four conditions (i.e. black and white groups × C and NC tests). Hence any black–white differences in items' Δ' values indicate only differences in the *relative* item difficulty, that is, the difficulty of each item relative to the average level of difficulty *within* each set of items (C or NC)

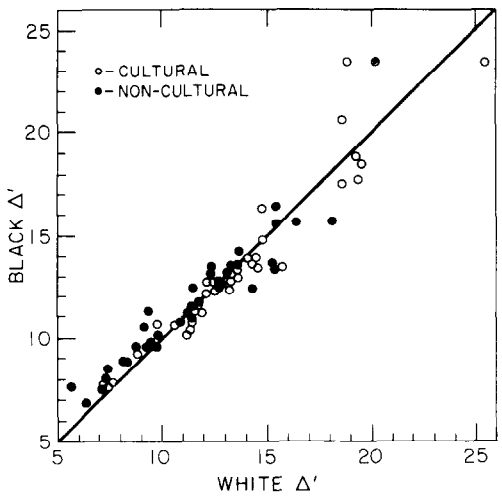

Fig. 1. Bivariate plot of the Δ′ (delta prime) index of relative item difficulty for the 37 C and 37 NC items in the black and white groups. The Δ′ values are scaled so as to have a mean of 13 and SD of 4 in each group and for both the C and NC items.

and *within* each group (black or white). A significant difference between blacks and whites in Δ′ for a given item indicates item bias, for whatever reason. Figure 1 shows a bivariate plot of the Δ′ values of the C and NC items for blacks and whites. If there were a complete absence of item biases of any kind, all of the points would fall exactly on the diagonal line. Departures from the diagonal indicate some degree of item bias. Points that fall above the diagonal indicate items that are relatively more difficult for blacks than for whites, and the points below the diagonal are items that are relatively more difficult for whites. Overall, there are no significant differences in the number of item biases favoring either racial group on either type of items, as indicated by the following contingency table, which yields $\chi^2 = 0.49$, $df = 1$, $P = 0.48$.

	C	NC	
Relatively more difficult for blacks	18	15	33
Relatively more difficult for whites	19	22	41
	37	37	74

The black–white difference in Δ′ values for single items can be tested for significance by

$$\chi^2 = (\Delta'_W - \Delta'_B)^2 / \mathrm{SE}^2_{\Delta\mathrm{diff}}, \text{ with 1 } df.$$

The standard error of the difference between Δ′ values is

$$\mathrm{SE}_{\Delta\mathrm{diff}} = \sqrt{[16/(N_W - 1) + 16/(N_B - 1)]}$$

[Note: This formula is *incorrect* as given in Jensen (1980), p. 461, formula 9N.20.] By this test of significance, 21 of the total of 74 items show black–white Δ′ differences significant beyond the 0.05 level of confidence, distributed as follows:

	Items biased against		Not significantly biased
	Blacks	Whites	
C items	5	6	26
NC items	5	5	27

Although, according to this criterion, there are significant item biases in both the C and NC tests, the direction of the significant biases are rather evenly balanced between blacks and whites and do not differ between C and NC items.

Hence the greater mean black–white difference on the NC test than on the C test cannot be attributed to item bias as assessed by any of the methods applied.

Characterization of C and NC items

Ideally, we would factor analyze all of the C and NC items together, including each item's correlation with race (quantitized as 0 and 1), and examine the nature of the factors on which race had appreciable positive or negative loadings. A factor analysis was attempted, using both phi coefficients and tetrachoric correlations between items, but in each case the correlation matrix was 'ill-conditioned', or non-Gramian, probably because of the wide range of item variances. A worthy factor analysis at the item level unfortunately was not obtainable.

The C and NC items, however, were classified by inspection according to item types, with the following result:

Cultural items	Noncultural items
Interpretation of common proverbs	Spatial visualization
Reading comprehension	Figural analogies
Incomplete sentence	Verbal analogies
Vocabulary, word knowledge	Verbal opposites
Algebra-type problems	Arithmetic problem solving
	Letter series completion
	Number series completion
	Syllogisms
	Clock problems
	Alphabetic sequence problems

In general, the C items are more dependent on information gained by subjects prior to taking the test. The NC items, on the other hand, contain all the information required for solution within the item itself, so that achieving the correct answer depends upon properly manipulating the given information of figuring out the solution from the essentially simple and familiar information provided in the item. This distinction between the recall of past-learned information and the mental manipulation of simple and familiar information that is provided in the test item itself is apparently the implicit basis on which McGurk's judges classified test items as being more or less culturally loaded.

Hence it turns out that it is not really anything anout the items that can be explained in terms of differences in cultural content or cultural item biases *per se* that accounts for the greater black–white difference on the NC than on the C items, but rather a difference in the degree of the items' requirements of reasoning and problem solving, as contrasted with recognition or recall of acquired information, for attaining the correct answer. Item biases associated with racial differences can probably be best understood, not in terms of presumed differences in degree of cultural loading of items, but in terms of the specific kinds of cognitive processes required by different items.

REFERENCES

Eysenck H. J. (1984) The effect of race on human abilities and mental test scores. In *Perspectives on Bias in Mental Testing* (Edited by Reynolds C. R. and Brown R. T.). Plenum, New York.

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

Jensen A. R. (1980) *Bias In Mental Testing*. Free Press, New York.

Jensen A. R. (1984) Test bias: concepts and criticisms. In *Perspectives on Bias in Mental Testing* (Edited by Reynolds C. R. and Brown R. T.). Plenum, New York.

McGurk F. C. J. (1951) *Comparison of the Performance of Negro and White High School Seniors on Cultural and Noncultural Psychological Test Questions*. Catholic University Press, Washington, D.C.

McGurk F. C. J. (1953a) On white and Negro test performance and socioeconomic factors. *J. abnorm. soc. Psychol.* **48**, 448–450.

McGurk F. C. J. (1953b) Socioeconomic status and culturally weighted test scores of negro subjects. *J. appl. Psychol.* **37**, 276–277.

McGurk F. C. J. (1967) The culture hypothesis and psychological tests. In *Race and Modern Science* (Edited by Kuttner R. E.), pp. 367–381. Social Science Press, New York.

McNemar Q. (1949) *Psychological Statistics*. John Wiley, New York.

Scarr S. and Saltzman L. (1982) Genetics and intelligence. In *Handbook of Human Intelligence* (Edited by Sternberg R. J.), pp. 792–896. Cambridge University Press, Cambridge.

Sims V. M. (1927) *The Measurement of Socioeconomic Status*. Public School Printing Co., Bloomington, IL.

Wigdor A. K. and Garner W. R. (Editors) (1982) *Ability Testing: Uses, Consequences and Controversies. Part 1: Report of the Committee; Part 2: Documentation Section*. National Academy Press, Washington, D.C.