

RACE, SOCIAL CLASS AND ABILITY PATTERNS ON THE WISC-R

ARTHUR R. JENSEN

Institute of Human Learning, University of California, Berkeley, CA 94720, U.S.A.

and

CECIL R. REYNOLDS

Department of Educational Psychology, 704 Harrington Education Center, Texas A & M University,
College Station, TX 77843, U.S.A.

(Received 5 January 1982)

Summary—The national standardization sample of whites and blacks on the Wechsler Intelligence Scale for Children—Revised (WISC-R) was the basis for a detailed analysis of the psychometric nature of racial and social class differences on the original 13 subscales of the WISC-R. The profiles of subtest scores of whites and blacks were compared directly and also after the racial groups were statistically equated on Full Scale IQ (FSIQ). Under the latter condition, the races differ only very slightly, although significantly, on some of the subtests, in ways generally contrary to popular expectations. The profile of white-black differences on the WISC-R subtests is markedly different, and negatively correlated with, the profiles of social class differences *within* each racial group, indicating that the pattern of racial differences is not explainable in terms of the difference in the average socio-economic status (SES) of blacks and whites. A Schmid-Leiman orthogonalized hierarchical factor analysis yields virtually identical factor structures and highly congruent factor loadings on the subtests for whites and blacks. Analysis of factor scores shows that by far the largest proportion of the variance between races is attributable to the general factor (*g*) common to all the subtests, whereas the group factors (verbal, performance and memory) contribute only minutely to the interracial variance. Hence the white-black differences on the diverse subtests of the WISC-R, and in the Full Scale IQ, are interpreted primarily as a difference in Spearman's *g*, rather than as differences in the more specific factors peculiar to particular content, knowledge, acquired skills or type of test. However, some slight but significant differences in patterns of ability also occur that are independent of *g*.

INTRODUCTION

The disparity of about one standard deviation, equivalent to 15 or 16 IQ points, between the mean scores of whites and blacks on tests of intelligence standardized on representative samples of the United States population is a well established fact in differential psychology (e.g. see Reynolds and Gutkin, 1981; and Reynolds and Brown, 1982). A logical and necessary first step toward understanding this phenomenon, and one which must precede the formulation of any causal hypotheses, is to describe precisely the nature of the white-black difference reflected by the overall scores on standard tests. Specifically, in which abilities, factors, contents, or formal features of such tests do whites and blacks differ the most and the least? Are the racial differences homogeneous for all types of items, or for all of the ability factors that contribute to the total variance? Or are the racial differences heterogeneous, so their directions and magnitudes can be more precisely specified as to test characteristics?

Past attempts to answer such questions have generally compared different racial or cultural groups on a variety of tests that presumably measure a number of different abilities. Standardized scores on each test are obtained for the combined groups, and then the profiles of mean scores on the various tests are obtained separately for each group. Often the groups show quite different profiles, as in the classic study by Lesser *et al.* (1965), in which Chinese, Jewish, Black and Puerto Rican children were compared on tests of verbal, reasoning, number and spatial abilities. (The few existing other similar studies have been reviewed by Willerman (1979).) A major problem with these studies,

spelled out in detail by Jensen (1980, pp. 729–736), is that tests of various aptitudes, such as those used by Lesser *et al.* (1965) do not represent independent factors of ability. Each test may measure some abilities common to one or more of the others, so the particular profile of abilities observed for any given group is merely a rather fortuitous artifact of the way the tests were chosen or constructed. It is a common finding, for example, that differential aptitude tests are each usually more highly loaded on the general factor common to all of the tests than on the particular group factor for which the test is named, e.g. verbal, numerical, spatial etc. Thus, even if the subpopulations of interest differed only on the general factor, and if the various aptitude tests were differentially loaded on the general factor, it would superficially appear that the subpopulations differ more on some aptitudes and less on others. But this could be entirely an artifact of the various tests' different loadings on the *g* factor (or any other group factor common to two or more of the tests), and it could be possible that the subpopulations actually do not differ at all on any of the specific aptitudes measured by the several aptitude tests. Moreover, unless it is established that the various tests in fact measure the same aptitudes, and measure them to the same degree in the subpopulations of interest, comparisons of the subpopulations on the various test scores are psychologically meaningless. These ambiguities as to what exactly the groups differ on can be answered only by multivariate techniques, such as factor analysis and multiple regression analysis, that permit contrasting the groups on independent (i.e. uncorrelated) sources of variance in subtest scores. We have adopted these techniques for the present study, to analyze the white–black difference on the currently most widely used individual test of intelligence, the Wechsler Intelligence Scale for Children—Revised (WISC-R; Wechsler, 1974).

This type of multivariate analysis of white–black differences on WISC-R subtests is intended to throw light on several disputed issues when the tests are used and the scores interpreted in the same way for both racial groups. Does the racial difference reflect a rather uniform general difference common to all the items and subtests or does it reside in relatively specific contents, item types, or factors in the test, as one might expect, for example, if the various subtests had differing amounts of cultural bias? Do the WISC-R subtests measure the same ability factors for blacks and whites? More specifically, does the WISC-R have the same factor structure, and are the subtests similarly loaded on the same factors, in the two racial groups? Is the pattern of subtest differences between blacks and whites the same as the pattern of subtest differences between higher and lower SES groups within each race, in accord with the hypothesis that the average white–black difference in test scores merely reflects the lower socioeconomic status of blacks? All these questions arose in the course of two recent court trials, *Larry P.* (in California) and *P.A.S.E. vs Hannon* (in Chicago), in which racial–cultural bias in the WISC-R was contested, resulting in opposite decisions by the courts in the two cases. But in neither trial were these questions about white–black differences in the WISC-R answered in terms of properly objective, empirical evidence and statistical analyses specifically aimed to answer these questions. The present study should help to fill this void.

METHOD

Subjects

The data of this study are taken from the national standardization sample for the Wechsler Intelligence Scale for Children—Revised (WISC-R), which consists of 2200 children between the ages of 6 and 16½ yr, with approximately equal numbers of both sexes. The children were chosen by a stratified, random sampling procedure to be representative of the total population of the United States, based on the demographic features revealed in the 1970 census. The sample was stratified on the basis of age, sex, race, SES status, geographic region of residence, and urban vs rural residence. The sample contained 305 blacks and 1870 whites. The present analyses, however, include only 1868 whites, as 2 were dropped because one or another item of the test data or demographic data were incomplete. These WISC-R standardization data have been de-

scribed in greater detail elsewhere (Kaufman and Doppelt, 1976; Reynolds and Gutkin, 1979; Wechsler, 1974).

Procedure

The various WISC-R subtests are listed by their familiar names in Table 1. One of the 13 subtests—Tapping Span—is probably unfamiliar to most users of the WISC-R. It was included in the original standardization but was not included in the final published version of the WISC-R, mainly because it had the lowest correlation of any subtest with all the other subtests and hence contributed the least of any of the subtests to the total variance in Full Scale IQ. For the purposes of our analysis, it is desirable to retain the Tapping test in the battery. The Tapping test is an adaptation of the Knox Cubes test (Knox, 1914). It is primarily a nonverbal test of visual imitative memory. A straight row of four 1" wooden blocks, spaced 1" apart, is placed before the S. The examiner, holding a 1" block between his thumb and index finger taps out a pattern on the row of four blocks, say 1-4-2-3, if we imagine the blocks are consecutively numbered from left to right. The S's task is to immediately imitate the same pattern of taps on the row of blocks, tapping out the pattern just as the examiner had done. Task difficulty is increased by tapping out longer and more complex series.

The use of subtest scaled scores, with a mean of 10 and standard deviation of 3 at every age, in effect obviates age differences in the test scores and permits analyses to be based on the entire standardization sample, with its statistical advantage of a very large N . Hence all the analyses in this study are based on the age-standardized scaled scores.

A variety of statistical techniques was used for comparing and contrasting the psychometric features of WISC-R performance in terms of race (white and black) and SES. These techniques are best explicated in connection with the results of each of the particular analyses summarized in the following section.

RESULTS AND DISCUSSION

Profile of white-black differences on WISC-R subtests

Table 1 shows the means and SDs of the white and black samples on each of the WISC-R subtests. The mean differences are expressed in two ways: (1) the difference

Table 1. Mean and SD of scaled score ($\mu = 10, \sigma = 3$) of WISC-R subtests and Verbal, Performance and FSIQ ($\mu = 100, \sigma = 15$) for whites ($N = 1868$) and blacks ($N = 305$)

WISC-R Scale	White		Black		Difference*	
	Mean	SD	Mean	SD	$\bar{W}-\bar{B}$	Diff./ \bar{s}
Information	10.41	2.91	8.09	2.65	2.32	0.81
Similarities	10.29	3.01	7.91	2.92	2.39	0.79
Arithmetic	10.37	2.84	8.63	2.75	1.74	0.61
Vocabulary	10.42	2.94	7.86	2.76	2.56	0.88
Comprehension	10.44	2.81	7.83	2.53	2.61	0.94
Digit Span	10.08	3.00	9.18	3.19	0.90	0.31
Tapping Span	10.09	2.87	9.12	2.95	0.97	0.33
Picture Completion	10.41	2.87	8.12	3.03	2.29	0.79
Picture Arrangement	10.37	2.91	8.10	3.03	2.27	0.77
Block Design	10.39	2.92	7.70	2.70	2.69	0.93
Object Assembly	10.73	3.01	7.89	2.96	2.47	0.82
Coding	10.22	3.30	8.86	2.93	1.35	0.47
Mazes	10.41	3.06	8.39	3.22	2.01	0.69
Verbal IQ	102.02	14.19	87.82	13.15	14.20	1.01
Performance IQ	102.17	14.14	87.17	13.35	14.03	1.07
Full Scale IQ	102.25	14.08	86.42	12.75	13.90	1.14

* $\bar{W}-\bar{B}$ is the difference in mean scaled scores of whites and blacks: Diff./ \bar{s} is the mean difference divided by the weighted average SD of the white and black samples, i.e.

$$\bar{s} = \sqrt{(N_w s_w^2 + N_b s_b^2)/(N_w + N_b)},$$

where W and B stand for white and black, N is sample size, and s is the SD.

between the scaled scores, which are based on the entire standardization sample, in which the subtest scores are scaled to a mean of 10 and a SD of 3, whereas the Verbal, Performance and Full Scale IQ (FSIQ) are scaled to a mean of 100 and SD of 15; and (2) the mean scaled score difference divided by the average within-groups SD, symbolized by \bar{s} . The \bar{s} scale has the advantage of expressing group differences on all tests in a common unit, i.e. the mean within-group SD, as explained in the footnote to Table 1. If all we are interested in is the profile of differences between the racial samples on the various subtests, it matters little on which scale the differences are expressed: the correlation (Pearson r) between $\bar{W}-\bar{B}$ and $\text{Diff.}/\bar{s}$ for the 13 subscales in Table 1 is 0.996. This correlation is an index of profile similarity, which in this case is seen to be very high indeed.

Differences in subtest reliability can affect the profile of mean differences, and therefore should be taken into account. The mean difference in \bar{s} units can be corrected for attenuation by dividing it by the square root of the test's reliability coefficient. This correction was applied to the white-black differences in subtest means (in \bar{s} units), using the subtest reliabilities based on the entire standardization sample. Although the differences are all slightly increased by the correction for attenuation, its effect on the profile of the differences across subtests, which is our main interest here, is negligible. The correlation between the corrected and uncorrected profiles is 0.990. Therefore, there is no need to correct for attenuation in our further analyses of the profile of white-black differences. It could make no significant difference in the results.

Profile of white-black differences in subtests with FSIQ held constant

Another way of expressing the difference between two groups is by means of the point-biserial correlation. In this case, the point-biserial correlation expresses the degree of relationship between standard scores on each subtest and the S 's racial classification, with black and white quantized as 0 and 1, respectively. A positive correlation coefficient, therefore, indicates that whites score higher than blacks, on average. In the range of group differences considered in this study, the relationship of group mean differences to point-biserial correlations is almost perfectly linear. This means that the profile of white-black differences across the 13 subtests is virtually identical when expressed in terms of mean $W-B$ differences on the subscales or as point-biserial correlations between each subscale and the racial dichotomy.

The point-biserial correlation is used in the following analyses because, by means of partial correlations, we can view the profile of white-black differences when the FSIQ is statistically held constant. This allows an answer to the question to what extent do whites and blacks differ on the various WISC-R subtests when the two groups are statistically equated on overall level of intelligence as measured by the FSIQ.

The upper profile in Fig. 1 shows the zero-order point-biserial correlations for the 13 subscales, as well as for the FSIQ. (These zero-order point-biserial r s are correlated +0.993 with the actual mean white-black differences in scaled scores.) The lower profile shows the partial point-biserial r s, with FSIQ partialled out. Partial r s significantly greater than 0 beyond the 0.05 level of confidence are indicated by asterisks. The mean of the absolute values (i.e. irrespective of sign) of the partial r over all 13 subtests is 0.04, $SD = 0.02$, as compared with a mean of 0.24, $SD = 0.07$ for the 13 zero-order correlations (i.e. the upper profile in Fig. 1). In SD units, the mean point-biserial r of 0.24 corresponds to 0.70σ ; the mean partial r of 0.04 corresponds to the relatively small difference of approx. 0.12σ .

We see that whites and blacks statistically equated on IQ differ significantly, in favor of whites, on Block Design, Object Assembly, Comprehension and Mazes, and also differ significantly, in favor of blacks, on Arithmetic and Digit Span. It should be kept in mind that these are the significant white-black differences in subtests that remain after the overall group difference in general ability, as represented by the FSIQ, is statistically removed. In view of the popular belief that the Vocabulary subtest is particularly biased against blacks, it is noteworthy that the partial r for this subtest is exactly zero. In

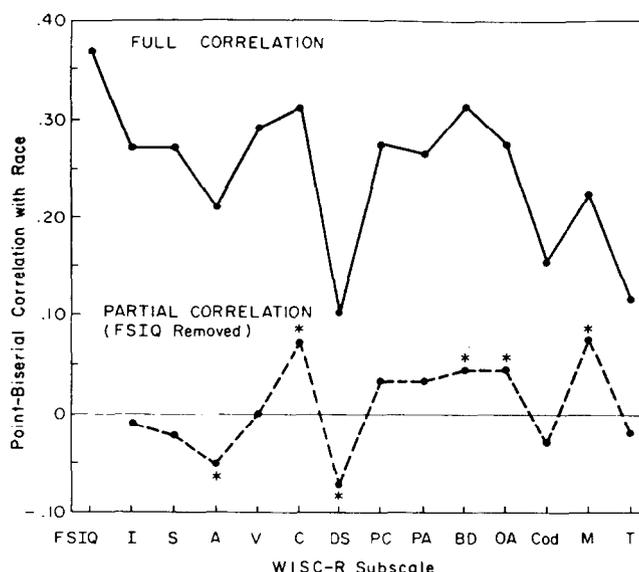


Fig. 1. Point-biserial correlation as an index of white-black mean difference on FSIQ and on each of 13 subtests of the WISC-R (Wechsler Intelligence Scale for Children—Revised). The upper profile shows the actual group differences. (All are statistically significant.) The lower profile shows the white-black differences on the 13 subtests after FSIQ has been partialled out, in effect equating the racial groups on general intelligence. Those differences which are significant beyond the 0.05 level are indicated by asterisks. I, Information; C, Comprehension; A, Arithmetic; S, Similarities; DS, Digit Span; V, Vocabulary; Cd, Coding [Digit Symbol]; PC, Picture Completion; BD, Block Designs; PA, Picture Arrangement; OA, Object Assembly; M, Mazes; T, Tapping [Knox Cubes].

general, it appears that the largest differences, independent of FSIQ, favoring whites are the performance tests that involve some spatial-perceptual ability, such as Block Design, Object Assembly and Mazes, whereas the tests favoring blacks, independently of FSIQ, are those involving short-term memory, e.g. Arithmetic and especially Digit Span. These two subtests, which patently involve short-term retention and retrieval of information, have also been characterized as reflecting 'freedom from distraction' due to anxiety in the testing situation. If this interpretation of Arithmetic and Digit Span is employed, these results are in direct contradiction to the anxiety hypothesis of black-white score differences (e.g. see Vernon, 1979; Reynolds and Jensen, 1980).

We can partition the total variance between races into two components: (1) that which is accounted for by a difference in general level of ability as indicated by FSIQ; and (2) that which is accounted for by each racial group's deviations from the general level on the 13 subtests, i.e. the profile of subtest differences between the groups. This is accomplished by comparing the point-biserial r^2 between race (white = 1, black = 0) and FSIQ with the squared multiple correlation, R^2 , between race, as the dependent variable, and the 13 subtests (in addition to FSIQ) as the independent variables. For all 13 tests plus FSIQ, the shrunken $R^2 = 0.170$; for FSIQ alone, $r^2 = 0.135$. Hence the subtest profile differences contribute $0.170 - 0.135 = 0.035$ to the maximum degree of discrimination ($R^2 = 0.170$) between the racial groups that can be attained with any additive combination of the WISC-R subscales. The FSIQ or general level of performance contributes 3.86 times as much to the maximum racial discrimination as the variation in subtest profiles.

Profile of SES correlations with WISC-R subtests

The white-black differences on psychometric tests are often attributed to the fact that the two populations differ in SES. If the racial difference in SES were an adequate explanation of the observed racial differences on various psychometric tests, we should expect to find essentially the same profile of SES differences as of race differences on the

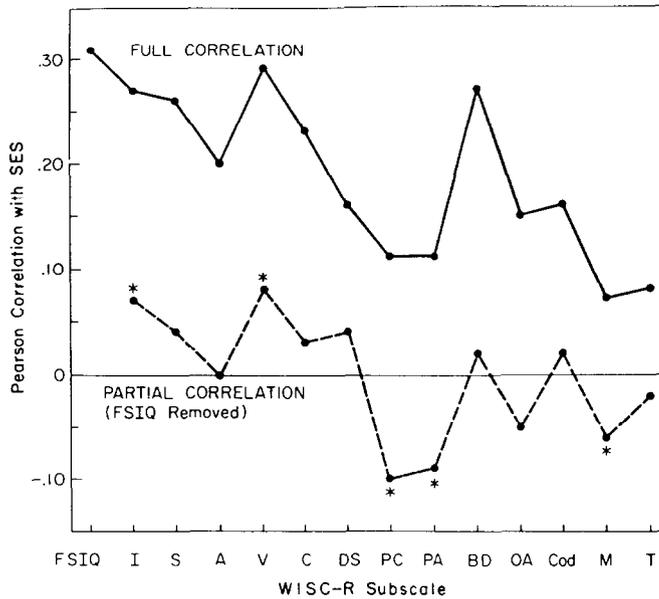


Fig. 2. Pearson correlations (r) between SES and scaled scores on FSIQ and on each of the 13 subtests of the WISC-R in the white sample ($N = 1868$). The upper profile shows the full (i.e. zero-order) correlations. (All are statistically significant beyond the 0.01 level.) The lower profile shows the partial correlations, with FSIQ partialled out. Differences significant beyond the 0.05 level are indicated by asterisks. (The subscales are fully identified in Fig. 1.)

13 subtests of the WISC-R. We can test this hypothesis, since the entire WISC-R standardization sample was classified into five conventional SES categories based on parents' occupation. On a 5-point scale of SES, the white mean is 2.27, $SD = 1.15$; the black mean is 1.49, $SD = 1.17$. The SES difference in \bar{s} units is 0.67.

The upper profile in Fig. 2 shows the zero-order correlations in the white sample between each of the WISC-R scales and SES; the lower profile shows the partial correlation, with FSIQ removed. Figure 3 shows the parallel results in the black sample. It should be noted that methodologically Figs 2 and 3 are directly comparable to Fig. 1 as regards the form of the profiles. For the time being we shall leave aside the question of comparing the overall relative magnitudes of race and SES in the composition of the total WISC-R variance. At this point, our interest is only in comparing the race and SES profiles.

The correlations of SES with each of the subtests, like the race correlations, are greatly reduced by partialling out FSIQ. Both the SES differences and the race difference are largely differences in general level of ability. But when the difference in general ability is held constant by partialling out FSIQ, how similar are the profiles of the partial correlations of the subtests with SES to the corresponding profile of partial correlations of the subtests with race? A familiar index of similarity between profiles is the Pearson r , i.e. the correlation coefficient between the 13 pairs of corresponding values making up the two profiles being compared. Identical profiles would have a correlation of 1: profiles no more similar than chance would be correlated zero. Here are the correlations between the subtest profiles based on partial correlations (with FSIQ removed) in Figs 1, 2 and 3:

$$\text{Race (Fig. 1)} \times \text{White SES (Fig. 2): } r = -0.45$$

$$\text{Race (Fig. 1)} \times \text{Black SES (Fig. 3): } r = -0.63.$$

It is noteworthy that both correlations are negative. That is to say, the pattern of subtest differences between whites and blacks is quite different—almost the opposite—from the pattern of subtest differences associated with SES, and this is true for SES within the white and black groups separately. This result is a direct contradiction of the hypothesis

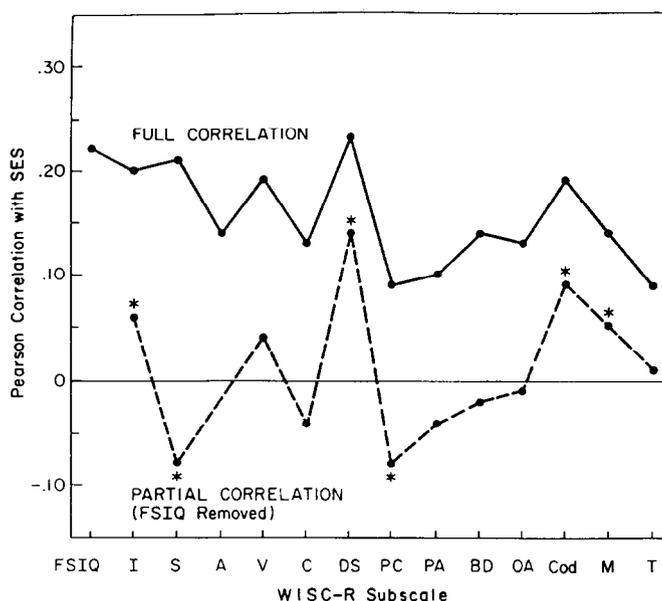


Fig. 3. Pearson correlations (r) between SES and scaled scores on FSIQ and on each of the 13 subtests of the WISC-R in the black sample ($N = 305$). The upper profile shows the full (i.e. zero-order) correlations. (All are statistically significant beyond the 0.05 level except PC and PA.) The lower profile shows the partial correlations, with FSIQ partialled out. Differences significant beyond the 0.05 level are indicated by asterisks. (The subscales are fully identified in Fig. 1.)

that the white-black difference in abilities reflects primarily the SES difference between whites and blacks. Whereas the white-black differences (independent of FSIQ) are largest on the spatial-perceptual subtests (in favor of whites) and on subtests involving short-term memory (in favor of blacks), the SES differences show up mainly on verbal ability (in favor of high SES) and spatial-perceptual ability (in favor of low SES). It should be noted, however, that the profiles of partial correlations between subtests and SES (i.e. Figs 2 and 3) are only moderately similar for blacks and whites, as indicated by a Pearson r of +0.59 between the profiles.

The apportionment of the total WISC-R variance associated with SES, *within* racial groups, to general level (FSIQ) and to the residual subtest profile differences is as follows:

<i>Whites</i>		
FSIQ:		$r^2 = 0.094$
Profile deviations:		$R^2 = 0.021$
<i>Blacks</i>		
FSIQ:		$r^2 = 0.049$
Profile deviations:		$R^2 = 0.016$

We note that SES is more highly correlated with WISC-R scores in the white than in the black sample. This does not seem to be entirely due to any statistical artifact, as the variances of whites and blacks on SES and on the WISC-R scales, and their reliabilities, are not sufficiently unequal to account for the marked differences in correlations. At present one can only speculate about the probable reasons that IQ is more highly correlated with SES in the white than in the black population.

Factor structure of WISC-R subtests in white and black samples

In this section we examine the degree of similarity between the white and black samples in the pattern of intercorrelations among the 13 WISC-R subtests and their factor structure.

Table 2. Intercorrelations (decimals omitted) among WISC-R subtests for whites (above diagonal) and blacks (below diagonal)

Subtest	I	S	A	V	C	DS	TS	PC	PA	BD	OA	Cod	M
Information		58	51	66	51	34	25	35	37	44	34	26	22
Similarities	55		43	63	55	33	19	40	37	45	35	25	24
Arithmetic	53	46		48	40	42	32	30	26	41	23	29	24
Vocabulary	63	65	52		61	36	24	38	39	43	33	29	21
Comprehension	49	48	39	63		23	19	35	34	38	29	23	23
Digit Span	43	34	50	41	35		37	16	18	29	17	28	18
Tapping Span	32	21	30	25	24	43		16	19	27	15	25	19
Picture Completion	42	43	32	43	44	28	29		34	47	41	15	29
Picture Arrangement	29	36	23	36	38	30	26	37		41	37	22	27
Block Design	37	41	40	41	38	35	26	48	37		56	30	39
Object Assembly	31	36	28	34	35	25	17	49	41	57		20	31
Coding	21	26	28	28	26	25	25	16	21	43	39		18
Mazes	26	24	22	25	30	28	26	36	32	29	19	18	

Table 2 shows the intercorrelations among the subtests for whites and blacks. The central tendency and dispersions of correlations is highly similar in the two groups. The mean r for whites is $+0.33$, $SD = 0.12$, and for blacks it is $+0.35$, $SD = 0.11$. The correlation (Pearson r) between the 78 paired analogous r s in the white and black matrices is $+0.89$, which indicates a high degree of similarity in the pattern of subtest intercorrelations for whites and blacks. [Despite the visibly high degree of similarity between the white and black correlation matrices, however, a sensitive statistical test (Jennrich, 1970) of the overall difference between the two matrices shows it to be significant ($\chi^2 = 102.21$, $df = 78$, $P = 0.041$). With the very large sample sizes of the present study, even quite minute differences are statistically significant.] As a consequence of this high degree of similarity between the white and black correlation matrices, we should also expect to find highly similar factor structures in the two groups, which has indeed been the major result in studies using different methodologies (Gutkin and Reynolds, 1981; Reynolds, 1982).

The correlations were subjected to a Schmid-Leiman (1957) hierarchical factor analysis, which seems the most appropriate for the factor analysis of ability tests.* This type of analysis begins with a principal factor analysis, with communalities in the principal diagonal. Three principal factors with eigenvalues greater than 1 were extracted. These factors were rotated obliquely to approximate simple structure. From the correlations among the three oblique (i.e. correlated) primary factors, which were here identified as verbal (V), performance (P), and short-term memory (M), was extracted a second-order general factor (g). In the WISC-R battery, this second-order factor is essentially the same as Spearman's g , which can be called 'general intelligence'. Finally, the g factor is, in effect, partialled out of the three primary factors, rendering them orthogonal (i.e. uncorrelated). Thus, the g factor and the three primary or group factors are all orthogonal. The results of this factor analysis are shown separately for whites and blacks in Table 3.

The percentage of the total variance accounted for by common factors is 45.3% for whites and 47.4% for blacks—a nonsignificant difference. We also note that the communalities (h^2) of the various subtests are highly similar for whites and blacks; the Pearson correlation between the 13 pairs of white-black h^2 values is $+0.91$. The percentage of the total variance accounted for by each factor is not significantly different for whites and blacks.

The degree of similarity between the groups on each factor is properly indicated by the Burt-Tucker coefficient of congruence, r_c (Cattell, 1978). This index is similar to the correlation coefficient, taking values ranging from 0 to ± 1 , with positive values indicating factorial similarity. The coefficient of congruence is calculated as follows:

* We are greatly indebted to John Schmid for kindly performing the Schmid-Leiman hierarchical factor analyses in this study. We are also grateful to Steven Paul and Emily Toy, who ably carried out all the other statistical analyses of these massive data.

Table 3. Schmid-Leiman hierarchical factor analysis of WISC-R subtests in white ($N = 1868$) and black ($N = 305$) samples

Subtest	White					Black				
	<i>g</i>	V	P	M	h^2	<i>g</i>	V	P	M	h^2
Information	67	35	01	09	58	65	35	-04	13	56
Similarities	67	36	06	00	46	62	41	07	-04	56
Arithmetic	57	16	00	32	45	60	21	-07	25	47
Vocabulary	72	46	-06	04	73	71	53	-03	-02	79
Comprehension	60	37	02	-03	50	61	32	11	-01	49
Digit Span	44	03	-05	49	44	59	00	-06	45	55
Tapping Span	35	-07	04	43	31	44	-09	01	38	35
Picture Completion	51	08	33	-08	38	57	11	34	-03	45
Picture Arrangement	49	09	25	-01	31	49	05	28	04	32
Block Design	65	-03	47	08	65	61	-01	42	05	55
Object Assembly	50	-03	47	-09	48	53	-02	54	-10	58
Coding	37	02	07	25	20	36	04	05	16	16
Mazes	37	-07	30	09	24	45	-10	34	10	34
% Variance	29.7	5.1	5.6	4.9		31.7	5.8	6.2	3.7	

Decimals omitted in factor loadings and communalities.

$$r_c = \frac{\sum b_1 b_2}{(\sum b_1^2 \sum b_2^2)^{1/2}},$$

where b_1 and b_2 are the factor loadings on a given test in groups 1 and 2. The congruence coefficients between whites and blacks on each of the four orthogonalized factors are as follows:

	r_c
General (<i>g</i>)	+0.995
Verbal (<i>V</i>)	+0.989
Performance (<i>P</i>)	+0.985
Memory (<i>M</i>)	+0.984

All of these congruence coefficients are of such magnitude as to represent virtual identity of the WISC-R factors in the white and black populations.

Factor scores and white-black differences

Estimated factor scores on the four factors, *g*, *V*, *P* and *M*, were obtained on each *S* in the combined samples ($N = 2173$). The factor scores were derived from the factor loadings in the white sample. Since the white sample is much larger than the black sample, the factor loadings in the white sample would have smaller sampling error. The very high congruence coefficients between the white and black factors suggest that the slight differences in factor loadings are practically negligible and no greater than would be expected from sampling error. Moreover, factor loadings based on a factor analysis of the correlations among the subtests within the combined samples risk being spuriously inflated because of the mean racial group differences on all of the subtests, and we do not wish to confound the magnitudes of the factor loadings with the group differences. The correlations between the factor scores and the Verbal, Performance, and FSIQs are shown in Table 4. All of the correlations are significant beyond the 0.001 level. It should be noted that both the Verbal IQ and the Performance IQ are more highly correlated with the *g* factor scores than with the factor scores on either the verbal or performance factors. Thus each of the WISC-R IQ scales measures *g* more than it measures any other factor. The FSIQ is a very close approximation to the *g* factor scores; the two are correlated +0.98.

Table 4. Correlation between factor scores and WISC-R IQ scales in the combined samples ($N = 2173$)

IQ scale	Factor			
	<i>g</i>	Verbal	Performance	Memory
Verbal	0.94	0.69	0.08	0.17
Performance	0.85	0.10	0.70	0.09
Full Scale	0.98	0.46	0.41	0.14

Although all the factors are themselves orthogonal (i.e. uncorrelated) in the Schmid-Leiman analysis, the factor scores derived in the combined groups are not uncorrelated, but show low to moderate intercorrelations (mean absolute $r = 0.32$, $SD = 0.11$) (see Table 6). There are two reasons for this: (1) the obtained factor scores are actually only estimates of the true uncorrelated factor scores; and (2) even if the estimated factor scores were perfectly uncorrelated *within* each racial sample, they could be correlated in the combined samples because of aggregation, i.e. there are overall group mean differences on all of the test variables that enter into the estimation of factor scores on each of the factors. Hence direct comparisons of the mean factor scores of the two racial groups on each of the factors would not show how much the groups differ on each of the four factors independently. To overcome this problem, we have performed a stepwise multiple regression analysis in which *race* (white = 1, black = 0) is the dependent variable and the four factor scores are the independent variables. The order of entering the factor scores in the stepwise analysis was determined solely by the size of the increment in R^2 contributed by each factor score. Our main interest is in seeing how much the group factors contribute to the racial discrimination *independently* of the *g* factor, and this type of analysis permits this determination. Each increment in the R^2 contributed independently by each factor can be expressed as a point-biserial correlation r_{pb} between the factor and the quantitized race dichotomy (white = 1, black = 0), and it can also be expressed as a mean difference between whites and blacks in units of the average standard deviation of factor scores within groups, \bar{s} . The r_{pb} and the mean difference in \bar{s} units on each of the four factors' successively independent discriminations between the races are shown in Table 5. All of the values of r_{pb} are significant beyond the 0.01 level, which means that the racial groups differ significantly on all four factors, although by far the largest difference is on the *g* factor. In terms of the total variance between races accounted for, the *g* factor accounts for more than seven times as much of the interracial variance as the other three factors (verbal, performance, memory) combined. It should be noted that blacks, on average, are superior on the memory factor whereas whites are superior on the *g*, verbal and performance factors. That is to say, when whites and blacks are equated on the *g* factor, blacks are superior on the short-term memory factor—about one-third of a standard deviation above whites. But the superiority of blacks on the memory factor has little

Table 5. The point-biserial correlation (r_{pb}) between the quantitized race variable (white = 1, black = 0) and independent factor scores, and the mean white-black difference

Factor	r_{pb}	r_{pb}^2	Diff./ \bar{s}
<i>g</i>	0.37	0.135	1.14
Memory	-0.10	0.010	-0.32
Verbal	0.06	0.004	0.20
Performance	0.06	0.004	0.20
Composite	0.39	0.153	1.22

Expressed in units of the average SD within groups, \bar{s} on the independent factor scores.

effect on the FSIQ, because the memory factor contributes such a small proportion of the total variance in this battery of tests. It is noteworthy that the mean white–black difference (in \bar{s} units) on the g factor is 1.137 \bar{s} , which is virtually the same as the mean white–black difference of 1.139 \bar{s} on the FSIQ.

Spearman's hypothesis of the white–black difference on psychometric tests

Spearman (1927, p. 379) conjectured, on the basis of a small amount of evidence, that the variable magnitudes of the mean difference between whites and blacks in standardized scores on a variety of mental tests are directly related to the size of the tests' loadings on g . Spearman did not attempt to test this hypothesis or to develop it any further than his original statement. The hypothesis can actually be stated in two forms, which can be termed *strong* and *weak*, respectively, although Spearman himself did not suggest this distinction. The strong form of the hypothesis is that the magnitudes of the mean white–black differences (in standard score units) on a variety of tests are directly related to the tests' g loadings because whites and blacks differ *only* on g and on no other cognitive factors. The weak form of the hypothesis is that the white–black difference in cognitive abilities is *predominantly* a difference in g , although the races might also differ, but to a much lesser degree, in certain other ability factors besides g .

The stepwise multiple regression analysis of factor scores presented in the preceding section clearly contradicts the strong form of Spearman's hypothesis, since each of the four factors extracted from the 13 WISC-R subtests was found to discriminate significantly ($P < 0.01$) between whites and blacks. The weak form of Spearman's hypothesis, however, is strongly borne out by the finding that the g factor accounts for more than seven times as much of the between-races variance as the other three factors combined, confirming the findings of Reynolds and Gutkin (1981) and Reynolds and Jensen (1980) with different methodological approaches to this question. This substantiation of the weak form of Spearman's hypothesis indicates that white–black differences in a variety of psychometric test scores are not essentially attributable to idiosyncratic cultural or linguistic peculiarities in this or that test, but to g , the general factor that all sorts of mental tests measure in common, and which some tests measure to a greater degree than others.

Spearman made his original conjecture after noting a parallel between white–black differences on a number of diverse tests and the tests' g loadings. We can make the same kind of observation in the present WISC-R subtests. In essence, it consists of calculating the correlation between the profile of mean white–black differences on the 13 WISC-R subtests, expressed in some form of standard units, with the profile of the tests' g loadings. The g loadings should be derived from each racial sample separately, rather than from the combined groups, to prevent any confounding between the racial group mean differences and the g factor loadings, which would cause spurious correlation between them.

Before proceeding with such an analysis, we should anticipate the predictable criticism that is sure to be made by someone, namely, that the hierarchical g factor that we have extracted by means of the Schmid–Leiman (1957) procedure (see Table 3) may not correspond well with the g factor that could be extracted by other possible methods. The three methods for estimating the general factor that are the most widely recognized today are, in order of our preference for theoretical reasons: (1) hierarchical factor analysis (with communalities in the principal diagonal); (2) principal factor analysis (communalities in principal diagonal); and (3) principal components (unities in the principal diagonal). In (2) and (3), the first unrotated principal factor and the first unrotated principal component are regarded as estimates of the g factor. It has been our empirical experience, in factor analyzing many different collections of ability tests, that the three methods listed above yield highly similar g factors. This finding is not a mathematical necessity of these methods, which could conceivably yield rather different g factors in some exceptionally unusual collection of tests that could be claimed to represent a peculiar sampling of the domain of psychometric abilities. In order to determine the degree of similarity of

g loadings extracted by the three methods, we have computed congruence coefficients among the three sets of g factor loadings separately for whites and blacks, as follows:

	White	Black
Hierarchical $g \times$ 1st Principal Factor:	+1.000	+0.999
Hierarchical $g \times$ 1st Principal Component:	+0.999	+0.999
1st Principal Factor \times 1st Principal Component:	+0.999	+0.999

Thus it is safe to say that the profile of g factor loadings of the WISC-R subtests is virtually the same regardless of the method of factor analysis used. This is equally true for whites and blacks.

Now we can look at the correlation between the profiles of black–white differences and of g loadings on the WISC-R subtests. In addition to Pearson r , we will also use Spearman's rank order correlation, ρ , as the index of similarity between the profile of white–black mean differences and the profile of g loadings over the 13 subtests, because ρ , unlike the Pearson r , is unaffected by possible noncomparability of the intervals throughout the scales of both variates. The use of Spearman's ρ , in addition to the Pearson r , is a check on possible distortion of our index of profile similarity by scale artifacts. If the values of the Pearson r and Spearman's ρ are about the same, we can be more confident of our assessment of profile similarity.

Comparison of the profile of point–biserial correlations (r_{pb}) as an index of mean white–black differences (see upper profile in Fig. 1) with the profiles of g loadings for blacks and whites (Table 3) show the following correlations:

	Pearson r	Spearman ρ
$r_{pb} \times g$ loadings (white)	+0.76	+0.75
$r_{pb} \times g$ loadings (black)	+0.56	+0.65

If we compare the profile of mean white–black differences in \bar{s} units (last column of Table 1) with the profiles of white and black g loadings (Table 3), we obtain the following correlations:

	Pearson r	Spearman ρ
Diff./ $\bar{s} \times g$ loadings (white)	+0.73	+0.75
Diff./ $\bar{s} \times g$ loadings (black)	+0.54	+0.64

Each of these correlations, with only 11 degrees-of-freedom, is significant beyond the 0.05 level. They are certainly consistent with the weak form of Spearman's hypothesis. The fact that these correlations are not higher than one might have expected, in view of the overwhelming importance of g as compared with other factors in the total between-groups variance in the WISC-R battery, is undoubtedly due in part to sampling errors both in the mean white–black differences and in the g loadings. There is also, unquestionably, some restriction of range of the g loadings relative to the domain of cognitive tests which could conceivably be sampled, because the test author chose the subtests on an *a-priori* basis out of a belief that they were good measures of g . Sampling errors would result in some variation in the relative values of the group differences and g loadings over the 13 subtests, and consequently the sampling error would attenuate the true correlation between differences and g loadings. To get an estimate of the sampling reliability of the profiles of group mean differences and of g loadings we used the split-half method. The white and black samples were each randomly divided into halves, and the profiles of white–black mean differences and g loadings were determined separately in each half. The average correlation between all six possible paired profiles of the four split-half white–black mean differences (in \bar{s} units) is 0.71. Because exactly the same mathematical

logic of the Spearman–Brown formula applies to random split-half samples with respect to the reliability of sample parameters as it applies to randomly split-half tests with the reliability of test scores, we can use the Spearman–Brown formula to boost the obtained split-half sample profile correlation of 0.71 to estimate the profile reliability for the entire sample. It is 0.83. Similarly, the split-half sample profiles of the g loadings are correlated 0.97 for whites and 0.87 for blacks; when boosted by the Spearman–Brown formula these are 0.98 and 0.93, respectively, and are estimates of the reliabilities of the profiles of g loadings in the total white and black samples. Using these reliabilities to correct for attenuation due to sampling error, the following (corrected) correlations are obtained:

	Pearson r
Diff./ $\bar{s} \times g$ loadings (white)	+0.81
Diff./ $\bar{s} \times g$ loadings (black)	+0.62

The results are not appreciably different for the profile of white–black differences expressed in terms of the profile of point–biserial correlations:

	Pearson r
$r_{pb} \times g$ loadings (white)	+0.84
$r_{pb} \times g$ loadings (black)	+0.64

The reason for the differences of about 0.20 between the above correlations for blacks and whites hardly seem worth speculating about in view of the fact that they have only 11 degrees-of-freedom and are about the size of their standard error. The correlations do indicate that Spearman’s original observation is largely substantiated by these WISC-R data. Although the correlation between white–black differences and g loadings is significant and substantial, it is not perfect because the races also differ, but to a much lesser degree, on other factors besides g .

Contribution of subtest specificities to the race difference

A test’s *specificity* is that proportion of its true score variance not accounted for by any of the factors it measures in common with any of the other tests that entered into the factor analysis. A test’s specificity can also be thought of as that aspect of whatever the test measures (other than error variance) that is uncorrelated with any other test in the battery. In the present hierarchical factor analysis of the WISC-R the percentage of the total variance attributable to specificity is approx. 50% for both whites and blacks. How much does the WISC-R specificity contribute to the total interracial variance? We can determine this by subtracting the unbiased (i.e. shrunken) multiple R^2 between the four factor scores (as the independent variables) and race (as the dependent variable) from the unbiased R^2 between FSIQ plus the 13 subtest scores (independent variables) and race (dependent variable). The resulting difference is the proportion of the total interracial variance attributable to the specificity of the subtests. This turns out to be $0.170 - 0.152 = 0.018$. Since the proportion of the total interracial variance on all of the WISC-R subtests is 0.170, the increment of 0.018 that the subtests contribute over and above the interracial variance of 0.152 on the four factors g , V , P and M , constitutes about 11% of the total interracial variance. The factor scores contribute 89% of the interracial variance. In other words, whites and blacks differ much more on the common factors of the WISC-R subtests than they differ on whatever is specific to the subtests.

It should be noticed that the overall interracial variance measured by R^2 is the result of weighting the racial difference on each subtest, regardless of the sign (i.e. direction) of the difference, so as to maximize the composite difference. As regards the direction of the subtest differences in terms of a more favorable score, the specificities of some of the subtests favor whites and some favor blacks. The fact that the specificities of each of the 13 subtests are *negatively* correlated (-0.56 for whites and -0.36 for blacks) with the

mean white-black differences (in \bar{s} units) on the subtests indicates that whatever is specific to each of the subtests tends, in general, to favor blacks. That is to say, the higher the tests' specificity, in general, the less is the white-black difference. It is almost the converse of Spearman's hypothesis, that the higher the tests' g loading, the larger is the white-black difference. It should be recalled that the proportion of the total interracial variance (R^2) accounted for by the FSIQ is only 0.135, as compared with R^2 of 0.152 for the four factor scores, and R^2 of 0.170 for the FSIQ plus all 13 subtest scores. FSIQ and g factor scores show very similar zero-order correlations with race: +0.368 and +0.367, respectively.

Race and SES components of variance in WISC-R factor scores

The aim of the following analysis is to reveal the proportions of the total variance in WISC-R factor scores associated with the variables of race and SES. It must be emphasized that no direct causal attribution is implied by this analysis, except for the fact that we know that SES is not a cause of racial variation. However, if either race or SES makes a significant *independent* contribution to the variance in ability, measured here as WISC-R factor scores, it necessarily means that the other variable (race or SES) cannot be the sole cause of the ability differences. Another caveat: *race* in this analysis implies anything and everything associated with the racial difference between whites and blacks in America; it does not have an exclusively biological or exclusively socio-cultural meaning. The *independent* contribution of race, with the contribution of SES statistically partialled out, refers to all aspects of the white-black difference *except* the difference in SES specifically assessed by the present index of SES, viz. parents' occupation rated on a 5-point scale ranging from professional-managerial to unskilled labor. The analysis does not logically permit inference concerning the relative proportions of the racial variance ascribable to different sources such as genetic and environmental. The same thing can be said of the SES variable in this analysis. We are not saying that analysis of direct causality, or of genetic and environmental components of racial and SES variation in abilities is not possible, in principle, but only that very special kinds of data are required for such an analysis, which would involve various types of kinship data and experimental designs in quantitative genetics. The present data meet none of these requirements. Nevertheless, it is psychometrically interesting to observe how race and SES are differentially correlated with the factor scores of the WISC-R.

We begin with all the zero-order correlations among the variables, shown in Table 6, as the basis for the multiple regression analysis shown in Table 7. The multiple correlation, R , was computed for each factor score (as the dependent variable), with race and SES as the independent variables. The order of entering the two independent variables is shown with race entered first and SES second in the stepwise regression, and then the reverse order is shown. R^2 indicates the proportion of total variance in the factor scores associated with the first demographic variable alone (step 1) and then with both demographic variables together (step 2). The value shown in parentheses is the proportion of variance contributed by the second variable independently of the variance attributed to

Table 6. Correlations (decimals omitted) among race (white = 1, black = 0), SES and WISC-R factor scores in total sample ($N = 2173$)

	Race	SES	g	V	P
SES	27				
g	37	36			
Verbal	15	21	46		
Performance	21	08	36	-45	
Memory	01	07	23	-25	-20

With 2172 *df.* correlations greater than ± 0.06 are significant beyond the 0.01 level.

Table 7. Multiple regression analysis of race and SES variance components of WISC-R factor scores in total sample ($N = 2173$)

Factor score	Order	R	R ²	Order	R	R ²	Partial r
<i>g</i>	1. Race	0.367	0.135	1. SES	0.357	0.127	Race (-SES) = +0.315
	2. SES	0.462	0.214 (0.079)	2. Race	0.462	0.214 (0.087)	SES (-Race) = +0.302
Verbal	1. Race	0.156	0.024	1. SES	0.215	0.046	Race (-SES) = +0.112
	2. SES	0.242	0.058 (0.034)	2. Race	0.242	0.058 (0.012)	SES (-Race) = +0.187
Performance	1. Race	0.217	0.047	1. SES	0.078	0.006	Race (-SES) = +0.205
	2. SES	0.219	0.048 (0.001)	2. Race	0.219	0.048 (0.042)	SES (-Race) = +0.030 (NS)*
Memory	1. Race	-0.014	0.000 (NS)*	1. SES	0.074	0.005	Race (-SES) = -0.032 (NS)*
	2. SES	0.080	0.006 (0.006)	2. Race	0.080	0.006 (0.001)	SES (-Race) = +0.079

* NS = nonsignificant at 0.05 level.

the variable entered first. The R^2 for the first-entered variable is, of course, an amalgam of both variables, due to the correlation between race and SES. The partial correlation in Table 7 is the correlation between each demographic variable and the WISC-R factor score, with the other demographic variable (-SES or -Race) statistically partialled out. The partial correlations, of course, tell essentially the same story as the R and R^2 analyses. Clearly, both race and SES independently are much more highly correlated with g than with any of the other ability factors. Also, on the ability factors other than g , race and SES part company quite distinctly. The partial correlations of race and SES with g factor scores are not significantly different ($t < 1$). But SES shows a significantly higher ($P < 0.02$) partial r with the Verbal factor scores than does race. Race shows a significantly ($P < 0.001$) higher partial r with the Performance factor scores than does SES. And Race and SES show significantly ($P < 0.001$) different (and even *opposite*) partial correlations with the Memory factor scores. As in the previously discussed contrasts between Fig. 1 and Figs 2 and 3, we see again, this time in terms of factor scores, that the *patterns* of race differences and SES differences vary quite distinctly. If the race differences in WISC-R factor scores merely reflected white-black differences in SES—a popular claim—it would seem impossible to explain the distinctly and significantly different *patterns* of correlations of the factor scores with race and SES, when race and SES are each statistically independent of the other in the correlations.

REFERENCES

- CATTELL R. B. (1978) *The Scientific Use of Factor Analysis in the Behavioral and Life Sciences*. Plenum Press, New York.
- GUTKIN T. B. and REYNOLDS C. R. (1981) Factorial similarity of the WISC-R for white and black children from the standardization sample. *J. educ. Psychol.* **73**, 227-231.
- JENNRICH R. I. (1970) An asymptotic χ^2 test for the equality of two correlation matrices. *J. Am. statist. Ass.* **65**, 904-912.
- JENSEN A. R. (1980) *Bias in Mental Testing*. The Free Press, New York.
- KAUFMAN A. S. and DOPPELT J. (1976) Analysis of WISC-R standardization variables in terms of the stratification variables. *Child Dev.* **47**, 165-171.
- KNOX H. A. (1914) A scale based on the work at Ellis Island for estimating mental defect. *J. Am. med. Ass.* **62**, 741-747.
- LESSER G. S., FIFER G. and CLARK D. H. (1965) Mental abilities of children from different social-class and cultural groups. *Mongr. Soc. Res. Child Dev.* **30**, No. 4.
- REYNOLDS C. R. (1982) The problem of bias in psychological assessment. In *The Handbook of School Psychology* (Edited by REYNOLDS C. R. and GUTKIN T. B.). Wiley, New York.
- REYNOLDS C. R. and BROWN R. T. (1982) Bias in mental testing: An introduction to the issues. In *Perspectives in Bias in Mental Testing* (Edited by REYNOLDS C. R. and BROWN R. T.). Plenum Press, New York.
- REYNOLDS C. R. and GUTKIN T. B. (1979) Predicting the premorbid intellectual status of children using demographic data. *Clin. Neuropsychol.* **1**, 36-38.

- REYNOLDS C. R. and GUTKIN T. B. (1981) A multivariate comparison of the intellectual performance of blacks and whites matched on four demographic variables. *Person. individ. Diff.* **2**, 175-180.
- REYNOLDS C. R. and JENSEN A. R. (1980) Patterns of intellectual abilities among blacks and whites matched on Full Scale IQ. Paper presented at the Annual Meeting of the American Psychological Association, Montreal, September.
- SCHMID J. and LEIMAN J. M. (1957) The development of hierarchical factor solutions. *Psychometrika* **22**, 53-61.
- SPEARMAN C. (1927) *The Abilities of Man*. Macmillan, New York.
- VERNON P. E. (1979) *Intelligence: Heredity and Environment*. Freeman, San Francisco.
- WECHSLER D. (1974) *Wechsler Intelligence Scale for Children-Revised*. Psychological Corporation, New York.
- WILLERMAN L. (1979) *The Psychology of Individual and Group Differences*. Freeman, San Francisco.