PERGAMON

Personality and Individual Differences 28 (2000) 593–599

PERSONALITY AND INDIVIDUAL DIFFERENCES

www.elsevier.com/locate/paid

Black–white differences on various psychometric tests: Spearman's hypothesis tested on American armed services veterans

Helmuth Nyborg[a,*], Arthur R. Jensen[b]

[a]*Institute of Psychology, University of Aarhus, DK 8240 Risskov, Denmark*
[b]*School of Education, University of California, Berkeley, CA 94720-1670, USA*

Received 17 February 1999; received in revised form 30 March 1999; accepted 26 April 1999

Abstract

Psychometric data (19 variables) on the cognitive abilities of large samples of American white (W) and black (B) male armed services veterans were factor analyzed to test Spearman's hypothesis that variation in the size of the mean W–B difference on various cognitive tests is directly related to variation in the size of the tests' loadings on the *g* factor. The hypothesis is strongly borne out by the data. Other factors independent of *g* showed no significant relationship to W–B differences in this battery of diverse tests. © 2000 Elsevier Science Ltd. All rights reserved.

1. Introduction

On inspecting a report of some early American data showing only the mean standardized differences between whites (W) and blacks (B) on a variety of mental ability tests, Spearman (1927, p. 379) noted that the W–B difference "was most marked in just those [tests] which are known to be most saturated with *g*". Spearman based this observation solely on his experience in factor analyzing a great many other tests in the British population; he himself never tested his conjecture empirically. In Spearman's day, therefore, his observation had only the status of a hypothesis. It has long been known, of course, that the *g* factor, derived from some type of

* Corresponding author. Tel.: +45-8942-4900; fax: +45-8942-4901.
E-mail address: helmurth@psy.au.dk (H. Nyborg)

PII: S0191-8869(99)00122-1

factor analysis of the correlations among a number of tests of mental abilities, is the one factor common to all of the tests that were entered into the factor analysis. It is also typically the largest factor, accounting for more of the total variance in all of the tests than any other factor and it often accounts for a larger proportion of the variance than all of the other significant factors combined.

Spearman's hypothesis has now been tested in 18 independent studies comprising many different psychometric tests and based on large samples. Spearman's hypothesis has been borne out in every study (Jensen, 1985, 1998, ch. 11), showing a highly significant and substantial positive correlation (averaging $+0.62$) between tests' *g* loadings and the magnitudes of the standardized W–B differences on those tests.

Why then still another test of Spearman's hypothesis? Mainly because it further extends the generality of the phenomenon to other samples of the W and B populations and extends the generalization to still another, rather unique, battery of cognitive tests. One criticism of some previous studies has been that, although they were based on independent subject samples, they used the same battery of tests, usually the Wechsler Scales, implying that the phenomenon first noted by Spearman might be just a characteristic of a particular collection of tests or of only certain types of cognitive tests. The particular collection of tests used in the present study is quite different from any psychometric batteries that have been used to examine Spearman's hypothesis. The importance and implications of the hypothesis (discussed at length by Jensen, 1998) warrants that it be definitively either proved or disproved.

2. Method

2.1. Subjects

The Centers for Disease Control (1988) provided an archival data set on 4462 males who had served in the United States Armed Forces. Approximately half of the sample had served in the Vietnam war. The original purpose in obtaining these data was to assess the long-term effects of these veterans' military service about 17 years after induction in the military. The total sample is fairly representative of the US population in race, education, income and occupations. However, it should be noted that a mandate of the US Congress prohibits all persons who score below the 10th percentile on a pre-induction general aptitude test from serving in the military. Therefore the lower tail of the distribution of ability is somewhat truncated in this sample, with a relatively stronger effect on the lower-scoring subgroups in the population, thereby possibly diminishing to some undetermined degree the differences between the subgroups' mean scores on some of the psychometric tests.

The two population subgroups selected from the total sample for the present study are all of the non-Hispanic whites (W) and African-Americans or blacks (B), for which there was complete psychometric data. The sample sizes are W ($N = 3535$); B ($N = 502$). The subjects' average age on entering the service in 1967 was 19.9 years (S.D. 1.7); the average age at which they were tested by the Centers for Disease Control was 37.4 years (S.D. 2.5). Both of the average age differences between the W and B samples are less than 2 months.

2.2. Psychometric variables

The test battery provides 19 experimentally independent variables that are highly diverse in the types of abilities, information content and cognitive skills called for. Five of the tests were administered at the time the subjects were inducted into the armed forces; all the others were administered approximately 17 years after induction, on average. The 19 test scores used in all of the analyses are briefly described as follows:

1. Grooved Pegboard Test (GPT), (right hand): a measure of manual dexterity and fine motor speed; the speed score is the reciprocal of the number of seconds taken to place a set of pegs in a grooved hole as quickly as possible.
2. GPT (left hand).
3. Paced Auditory Serial Addition Test (PASAT): a measure of mental control, mental speed and computational and attentional abilities. The subject mentally adds a sequence of numbers in rapid succession. score is the total number of correct responses.
4. Rey–Osterrieth Complex Figure Drawing (CFD), direct *copy* score: a measure of visual–spatial ability and memory; the subject reproduces a complex spatial figure while the figure is in full view.
5. CFD, copy from *immediate* recall.
6. CFD, copy from *delayed* recall (20 min of other activities intervening).
7. Wechsler Adult Intelligence Scale-Revised (WAIS-R), *general information*, scaled score.
8. WAIS-R, *block design*, scaled score.
9. Word List Generation Test (WLGT): a measure of verbal fluency; subject generates as many words as possible for 60 s that begin with each of three letters: F, A, S. Total number of words generated.
10. Wisconsin Card Sort Test (WCST): a measure of concept-formation, problem-solving and set-switching abilities and use of feedback in decision making. Ratio of correct responses to countable responses.
11. Wide Range Achievement Test (WRAT): measures ability to read aloud a list of single words (untimed). Total raw score.
12. California Verbal Learning Test (CVLT): a measure of verbal learning and memory; subject recalls a list of 16 words over five repeated learning trials. Total correct over 5 trials.
13. Army Classification Battery (ACB), *verbal test*, administered at time of induction. A measure of verbal reasoning.
14. ACB, *verbal test* administered an average of 17 years after induction.
15. ACB, *arithmetic reasoning test*, administered at time of induction.
16. ACB, *arithmetic reasoning test*, administered an average of 17 years after induction.
17. Pattern Analysis Test (PAT): a visual spatial measure of pattern recognition, administered at induction.
18. General Information Test (GIT): administered at time of induction.
19. Armed Forces Qualification Test (AFQT): a general aptitude battery; total score on four subtests (word knowledge, paragraph comprehension, arithmetic reasoning, mathematics knowledge). Administered at time of induction.

3. Results

Table 1 shows the effect size (*d*) of the W–B difference on each of the 19 variables and the first three principal components extracted from the correlations among the 19 variables in the total sample ($N = 4462$). The value of *d* for a given test is the mean W–B difference divided by the mean of the W and B standard deviations, which puts *d* on the same scale of S.D. units for all variables. The PCs are interpreted as: PC1, the general factor (*g*); PC2, visual–spatial memory and PC3, psychomotor speed and dexterity. We note that the same ACB tests administered about 17 years apart (indicated by *) have similar *d* values and PC loadings.

Principal components (PC) analysis, rather than any of the methods of common factor analysis, is used in this study, because we wish to test the W–B differences in uncorrelated factor scores and only PC analysis insures perfectly uncorrelated factor scores. (Factor scores derived from a common factor analysis are typically correlated to some degree, even though the factors themselves are perfectly uncorrelated with each other). Also, it has been shown that the first principal component (PC1) is about as good a measure of the general factor of a correlation matrix of ability tests as most forms of common factor analysis, such as principal

Table 1
Effect size (*d*) of the white–black difference and variables' loadings on the principal components with eigenvalues > 1 for 19 mental test scores in the total sample ($N = 4462$)

Variable	*d*	Principal components		
		PC1	PC2	PC3
(1) GPT, right hand	−0.402	0.363	0.363	0.734
(2) GPT, left hand	−0.474	0.371	0.379	0.712
(3) PASAT	0.726	0.599	0.039	0.099
(4) CFD, copy	0.529	0.510	0.415	−0.131
(5) CFD, immediate	0.619	0.582	0.611	−0.364
(6) CFD, delayed	0.611	0.581	0.615	−0.364
(7) WAIS-R, information	0.768	0.775	−0.259	−0.052
(8) WAIS-R, block design	1.164	0.697	0.281	−0.015
(9) WLGT	0.186	0.531	−0.200	0.076
(10) WCST	0.591	0.486	0.101	0.042
(11) WRAT	0.764	0.749	−0.349	0.053
(12) CVLT	0.462	0.519	0.027	−0.115
(13) ACB, verbal[a]	1.010	0.822	−0.346	0.010
(14) ACB, verbal	0.956	0.826	−0.318	0.033
(15) ACB, arithmetic[a]	1.148	0.817	−0.199	−0.024
(16) ACB, arithmetic	1.141	0.824	−0.119	−0.007
(17) PAT[a]	0.949	0.726	0.174	−0.55
(18) GIT[a]	1.284	0.710	−0.207	−0.032
(19) AFQT[a]	1.462	0.856	−0.041	−0.043
Eigenvalue		8.469	1.869	1.372
Percent of total variance		44.6	9.8	7.2

[a] Test administered at induction into the armed services. All other tests administered an average of 17 years after induction.

axis analysis or a Schmid–Leiman hierarchical factor analysis (Jensen & Weng, 1994). The latter is generally more desirable, but it does not yield uncorrelated factor scores and is therefore less appropriate for the present study. A Schmid–Leiman hierarchical factor analysis of these data shows a congruence coefficient of $+0.99$ between PC1 and the hierarchical *g*. Congruence coefficients of 0.95 and above are generally interpreted as indicating that the factors are virtually identical. The congruence coefficient closely approximates the Pearson correlation between the factor scores derived from the general factor represented by the PC1 and the hierarchical *g*.

Spearman's hypothesis is tested in two ways: (1) by the method of correlated vectors and (2) by comparing the mean W–B differences on factor scores derived from each of the significant factors (with eigenvalues > 1) in the correlation matrix for the 19 variables.

3.1. Correlated vectors

This method has been described in detail elsewhere (Jensen, 1998, Appendix B). It requires first that the same factors are measured in the W and B samples, so the PC analysis was performed separately in each sample. The preferred measure of factor similarity, based on the loadings of the tests on each factor (or PC), is the congruence coefficient. The coefficients of congruence between Ws and Bs for each of the PCs are: PC1 = 0.996, PC2 = 0.982 and PC3 = 0.975. Because a congruence coefficient of 0.95 indicates virtual identity of the factors for Ws and Bs, we can use the *averaged loadings* of the W and B groups in the subsequent analyses. (It would be incorrect to use the loadings in the combined samples, because these would also reflect the *between*-groups variance in addition to the *within*-groups variance, or individual differences among Ws and among Bs).

The method of correlated vectors consists of correlating the column vector of factor loadings with the column vector of the standardized size of the average W–B difference, *d*. As a check for outliers or any other peculiarities in the scale of factor loadings or of *d* that could spuriously bias the magnitude of the correlation, both the Pearson *r* and Spearman's rank-order correlation (r_S) were calculated. Close similarity between *r* and r_s assures against such a bias. The *t* test of significance of the correlation is based on r_s.

To control for the joint influence of differences in the various tests' reliability coefficients on both the factor loadings and the standardized W–B differences, a column vector of the tests' reliability coefficients is partialled out of the correlation between the vector of factor loadings and the vector of W–B differences, *d*. Because reliability coefficients of these tests have not been determined directly in a comparable subject sample, each test's communality (i.e. the proportion of its total variance accounted for by the common factors) is used as a lower-bound estimate of the test's reliability. Partialling out the vector of communalities (as surrogate reliability coefficients) is an extremely stringent procedure, because generally the largest proportion of the communalities is contributed by the PC1, so some part of the PC1 vector's correlation with *d* is removed in the partial correlation, thereby tending to work against the outcome predicted by Spearman's hypothesis. If the r_s is nonsignificant ($p > 0.10$, 2-tailed test), the partial correlation is not computed. Controlling variation in test reliabilities (estimated by the communalities), however, seems preferable to no control whatsoever. These results are shown in Table 2.

Table 2
Correlated vectors test of Spearman's hypothesis

Correlated vectors[a]	r	r_s	p	Partial r
PC1 × d	0.721	0.812	< 0.002	0.700
PC2 × d	−0.385	−0.423	n.s.	–
PC3 × d	−0.702	−0.232	n.s.	–

[a] Correlation based on number of variables in each set of correlated vectors = 19.

The regression of the d values on the PC1 loadings is: d = 1.46(PC1)− 0.07. Hence, theoretically a test that measured only the g factor (i.e. its PC1 loading = 1) would show a mean W–B difference of 1.39σ. This figure can be compared with the same determination based on 149 various psychometric tests administered to 15 independent samples (of various age groups, including children) comprising 43,892 Bs and 243,009 Ws (Jensen, 1998, p. 378). For these data the estimated W–B difference on a theoretical test that measured only g would be 1.31σ or 0.08σ (equivalent to 1.2 IQ points) less than the corresponding estimate of 1.39 in the present study.

3.2. Factor score differences

Factor scores (standardized to overall mean = 0, S.D. = 1) were obtained for every subject on each PC, The median and mean W–B differences and their significance levels are shown in Table 3, indicating a significant W–B difference only on PC1.

4. Discussion

The results clearly replicate the findings of previous studies of Spearman's hypothesis based on quite different sets of tests. The estimate of g in any reasonably large set of diverse cognitive tests is approximately the same g as in any other set of cognitive tests. Therefore, it should not be too surprising that various fairly representative samples of the W and B populations of the United States show the Spearman phenomenon. The present battery shows the effect, indicated by $r_s = 0.812$ (see Table 2), to a stronger degree than the average correlation of 0.62 for all previous studies (Jensen, 1998, ch. 11). The present study also shows

Table 3
W–B differences in median and mean of standardized PC factor scores (S.D. units) and the significance level (p) of the mean difference

Factor score	Median	Mean	p
PC1	+1.284	+1.174	< 0.00001
PC2	+0.024	+0.059	> 0.20
PC3	−0.002	−0.066	> 0.16

nonsignificant and negligible W–B differences in the factor scores derived from PC2 (spatial + memory) and PC3 (motor dexterity and speed). The W–B factor score differences on PC3, although it is nonsignificant ($p > 0.16$), has a negative value. However, there are only two motor skills measures defining this factor and these are virtually equivalent forms of one test (taken separately by the right and left hands). It is likely that a psychomotor factor determined by a larger number and variety of motor skill variables would show a considerably larger B–W difference (with Bs outscoring Ws) on the psychomotor factor scores. Despite its rather low positive loadings on *PC1*, the Pegboard motor-speed test shows an average negative W–B d value of −0.44, favoring Bs. A similar effect is seen in the Motor Coordination subtest of the General Aptitude Test Battery (GATB), in which Bs would score on average higher than Ws if the moderate g component of this test were removed statistically (Jensen, 1985). The same black advantage in motor speed also shows up in the faster movement speed of Bs than of Ws (and Asians) in speed-of-reaction tests that provide independent measures of decision speed and movement speed (Jensen, 1998, pp. 396–397). The test in the present battery with the smallest positive d value (0.186) and a relatively small PC1 loading (0.531) is the WLGT, a measure of verbal fluency.

The present findings, in addition to those of previous studies, which they strongly replicate, support the conclusion that Spearman's original conjecture that the W–B difference in cognitive tests is predominantly a difference in the g factor should no longer be regarded as just an hypothesis but as an empirically established fact.

References

Centers for Disease Control (1988). Health status of Vietnam veterans. *Journal of the American Medical Association*, *259*, 2701–2719.

Jensen, A. R. (1985). The nature of the black–white difference on various psychometric tests: Spearman's hypothesis. *Behavioral and Brain Sciences*, *8*, 193–219.

Jensen, A. R., & Weng, L.-J. (1994). What is a good *g*? *Intelligence*, *18*, 231–258.

Jensen, A. R. (1998). *The g factor*. Westport, CT: Praeger.

Spearman, C. E. (1927). *The abilities of man: their nature and measurement*. London: Macmillan.