Available online at www.sciencedirect.com
ScienceDirect
Intelligence 35 (2007) 47–57
INTELLIGENCE

Testing the cross-racial generality of Spearman's hypothesis in two samples

Peter Hartmann *, Nanna Hye Sun Kruuse, Helmuth Nyborg

Individual Differences Research Unit (IDRU), Department of Psychology, University of Aarhus, Nobelparken, Jens Christian Skous Vej 4, DK-8000 Aarhus C, Denmark

Received 30 August 2005; received in revised form 11 April 2006; accepted 12 April 2006
Available online 5 June 2006

Abstract

Spearman's hypothesis states that racial differences in IQ between Blacks (B) and Whites (W) are due primarily to differences in the *g* factor. This hypothesis is often confirmed, but it is less certain whether it generalizes to other races. We therefore tested its cross-racial generality by comparing American subjects of European descent (W) to American Hispanics (H) in two different databases. The first [Centers for Disease Control (1988). Health status of Vietnam veterans. *Journal of the American Medical Association 259*, 2701–2719; Centers for Disease Control (1989). *Health status of Vietnam veterans: Vol IV. Psychological and neuropsychological evaluation*. Atlanta, Georgia: Center for Environmental Health and Injury Control] contains 4462 middle-aged Armed Services Veterans males, and the second database (NLSY1979) holds 11,625 young male and female adults. Both samples are fairly representative of the general American population.

Race differences in general intelligence *g* were calculated and vectors of test scores were correlated with the vectors of the tests' *g* loadings, following Jensen [Jensen, A. R. (1998). *The g factor*. Westport, CT: Praeger].

W scored about 0.8 S.D. above H. The racial difference on the tests correlated significantly with the *g*-loadings of the tests in the VES database, but less so in the NLSY database.

We therefore conclude that the present study supports, but does not unequivocally verify, the cross-racial generality of the Spearman's hypothesis.

Keywords: Spearman's hypothesis; The *g* factor; Racial differences; Intelligence; Hispanic

1. Introduction

Spearman's hypothesis states that the difference in IQ scores between Whites (W) and Blacks (B) is larger on more *g* saturated tests (Spearman, 1927, pp. 379–380) and therefore primarily attributable to differences in *g* (Jensen, 1998), as compared to other cognitive factors such as stratum I and/or II factors (see Carroll, 1993).

The typical approach to testing Spearman's hypothesis is by categorizing subjects into different racial groups based on rated ethnicity (self or other) and/or on the basis of skin color. All subjects from all racial groups are then IQ tested and the group differences calculated in order to estimate the average differences in IQ. However, since IQ is an estimate of "intelligence in general" it may be contaminated with

* Corresponding author. Axel Heides Gade 12, 5.tv. 2300 Copenhagen S, Denmark. Tel.: +45 30600823.
E-mail address: peter.weber.hartmann@gmail.com (P. Hartmann).

0160-2896/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.intell.2006.04.004

stratum I and/or II factors. In order to separate *g* from these lower order factors, an extra statistical step is needed. The method of correlated vectors (see Jensen, 1998, chapter 11 and Appendix A) aims to accomplish this separation by correlating the vector of the subtests' *g* loading with the vector of the group differences on the subtests, while controlling for differences in reliability. The logic of this approach is that if a group difference in the test score is attributable to *g* then tests having a higher *g* loading should show a larger difference in test scores. The strength of such a relationship is reflected in the size of correlation coefficient between the two vectors. The current empirical understanding is that this hypothesis has often found confirmation, and the average correlation between the W–B differences and the *g* loadings of the cognitive ability tests amounts to 0.6 when using the method of correlated vectors (see Jensen, 1998, for review on this).

However, this method is not without its critics. Ashton and Lee (2005) have thus recently aired two important concerns about it. First, the presence of non-*g* variance can in fact cover up a relationship between the vector of *g* loadings and the vector of group differences in test scores, thereby in effect increasing the risk of committing a type II error. Second, the nature and content of the tests included in the analysis may affect the estimated *g* loading and then the subsequent correlation between the vectors. These objections suggest that the method of correlated vectors is not a bulletproof method for determining whether a group difference in IQ is attributable to a group difference in *g*. Despite these problems, we applied the method of correlated vectors in the present study simply because there is presently no blameless alternative.

The purpose of the present study is to go beyond the common W–B differences and test the cross-racial generality of Spearman's hypothesis for another race. To our knowledge the world literature witnesses only a dozen papers dealing with the cross-racial generality of Spearman's hypothesis.

Sandoval (1982) thus used the WISC-R (12 tests) to investigate the generality of Spearman's hypothesis for Americans of Anglo, African, and Mexican ancestry. The sample consisted of 953 subjects–332 Anglos, 314 Africans, and 307 Mexicans–all aged 5–11 years. Anglo subjects scored almost 1 S.D. above the Black and Mexican averages. Spearman's rank order correlations between the race differences and the subtest *g* loadings, calculated from the total as well as the Anglo sample, were obtained using the method of correlated vectors. When looking at the Anglo-Black difference, the rank order correlation was Rho = 0.48 ($p = 0.059$, two-tailed) when using the *g* loading estimated from the entire sample, but only Rho = 0.36 (n.s.) when using the *g* loading estimated from the Anglo sample alone. The result of the comparison of Anglo-Mexican was Rho = 0.78 ($p < 0.001$, two-tailed) and Rho = 0.82 ($p < 0.001$, two-tailed), respectively. Jensen (1998) rightfully cautions against using the *g* loadings of the total sample in correlated vector calculations, because this could affect the obtained correlation between the vector of *g* loadings and the vector of race differences. We accordingly find the Rho of 0.36 and 0.82 the best fitting estimates. Jensen (1998) also emphasizes the importance of correcting for unreliability. Actual corrected values were not presented, but the author stated that the correction only slightly increased the correlations. The study provided support for the cross-racial generality of Spearman's hypothesis with respect to Anglo-Mexican Americans, but only meager support for the original W–B Spearman hypothesis.

Nagoshi, Johnson, DeFries, Wilson, and Vandenberg (1984) used 15 cognitive ability tests in a correlated vector study of 6581 Hawaiian subjects of European, Japanese, and Chinese decent. Both parents and children participated. For parents, the vector of race differences in the subtests correlated between $r = 0.26$–0.32 (n.s.) with the *g* loading of the tests. For children, the correlations ranged from $r = 0.24$–0.58, and reached significance for the European–Chinese ($r = 0.51$; $p < 0.05$) and Japanese–Chinese populations (0.58; $p < 0.05$). In other words, across six comparisons only two reached significance. This provides only modest support for the cross-racial generality of the Spearman hypothesis.

Lynn and Owen (1994) subjected 1063 Indians and 1056 White subjects to the South African equivalent of the DAT (10 tests), and found an ability difference close to 1 S.D. in favor of White subjects. Correlated vector calculations reveal non-significant correlations ranging from $r = 0.081$–0.129 between the *g* saturation of the tests and the *g* loadings of the tests, depending on what sample was used for the estimation of the tests' *g* saturation. These observations challenge the cross-racial generality of Spearman's hypothesis with respect to White–Indian differences.

te Nijenhuis (1997) tested a Dutch majority group of 270 and a Dutch minority group of 247 young (both groups aged ca. 30) predominantly male subjects with a battery of *g* loaded safety tests. The tests measure selective attention, attentional speed, perceptual motor ability, sensory-motor coordination ability, and precision of reactions. As measure of the safety test *g*

loading, the correlation with GATB was used. The average ability difference was approximately 1 S.D. (measured in White majority S.D. units) in favor of the White majority subjects. A test for Spearman's hypothesis found Pearson correlations ranging from 0.59 to 0.81 and Spearman correlations from 0.59 to 0.61 (depending on whether the *g* loadings were disattenuated and what subgroup the *g* saturation was derived from). Although these correlations are moderate to large in magnitude, they did not reach significance (although they came close, $p<0.1$), and thus are only suggestive of the generality of Spearman's hypothesis.

te Nijenhuis and van der Flier (1997) compared a majority group of White subjects ($N=806$) in the Netherlands to four immigrant groups from Surinam ($N=535$), Dutch Antilles ($N=126$), North Africa ($N=167$), and Turkey ($N=275$). Using the GATB (8 tests), the authors found a close to 1 S.D. majority group lead (measured in majority S.D. units) in comparison to the minority groups.

The results of the correlated vector analyses are presented in Table 1.

Table 1 provides strong support for the cross-racial generality of Spearman's hypothesis about *g*-related IQ differences between native Whites and four immigrant groups in the Netherlands. However, the immigrant group from North Africa is arguably comparable to a "traditional Black sample", and thus provides support for the original Spearman's hypothesis rather than for the generality of Spearman's hypothesis.

te Nijenhuis, Evers, and Mur (2000) administered the Dutch version of the DAT (9 tests) to 318 White Dutch majority children and 111 unspecified Dutch minority children aged 12–13. The majority group averaged about 1 S.D. (as measured in majority S.D. units) in mean scores above the minority group. The result partly confirmed Spearman's hypothesis by finding a Pearson correlation of 0.78 ($p<0.05$) between the vector of *g* loadings and group differences. However, the Spearman rank order correlation (Rho = 0.60) did not reach significance although it came close ($p<0.1$). Furthermore, the authors also tested Spearman's hypothesis using school grades and two achievement tests (as opposed to ability tests) and found significant support for the hypothesis ($r=0.65$; Rho = 0.73).

Rushton (2002) reanalyzed data by Owen (1992) in order to test the Spearman hypothesis. The sample originated from South Africa and included 1056 White (W), 1063 Indian (I), 778 mixed-race Colored (C) and 1093 Blacks (B), all aged 14 and tested with Raven's Standard Progressive Matrices. The ability difference was measured in S.D. from the White mean and amounted to −1.35 for W–I, −0.52 for W–B, and −2.78 for the W–C difference. The Spearman rank order correlation between the racial difference in pass ratios and the items *g* loading amounted to Rho = 0.35–0.57 when using the *g* loading obtained from the White group and 0.61–0.85 when using the *g* loading from the respective minority group. All results were significant and hence confirm both the original and the generality of Spearman's hypothesis.

Rushton, Skuy, and Fridjohn (2002), administrated Raven's Standard Progressive Matrices to 342 17–23 years old South African engineering students, consisting of 198 Blacks, 86 Whites and 58 Indians. The obtained test scores corresponded to an IQ of 97, 110 and 102 (based on the 1993 US norm for 18–22 year olds). The Pearson and Spearman correlation between the racial differences in pass ratios and the item's *g* loadings (estimated through biserial and point-biserial correlation) were in general low and non-significant when using the *g* loading obtained from the White sample. When using the *g* loading obtained from the minority groups, the correlations were all significant and ranged from 0.26 to 0.67, depending on whether biserial or point-biserial correlations were used for estimating the *g* loading and whether Pearson or Spearman correlations were used for correlating the vectors. As for the W–B, the result was significant ($r=0.54$; Rho = 0.67) when using the point-biserial *g* loading, but not when using the biserial *g* loading ($r=0.00$; Rho = 0.16). Furthermore, the Jensen effect calculated for B–I indicated that, when using the point-biserial *g* loading obtained from the Black sample, the effect became significant ($r=0.45$; Rho = 0.38), but this was neither the case when using the biserial correlation nor when using the *g* loading obtained from the Indian group. The authors made a further analysis by aggregating items into 9–10 artificially created subtests, and then found all correlations to be either significant or close to significant ($p<0.1$), no matter which method was used for

Table 1
Correlated vector coefficients (compiled from te Nijenhuis & van der Flier, 1997)

Country of origin	Pearson's *r*	Spearman's Rho	Corrected Pearson's *r*
Surinam	0.72 *	0.42	0.76 *
Netherlands Antilles	0.77 *	0.71 *	0.78 *
North Africa	0.84 **	0.87 **	0.82 **
Turkey	0.70 *	0.87 **	0.64 *

* Significant at $p<0.05$.
** Significant at $p<0.01$.

obtaining *g* loading (biserial or point-biserial) or which type of correlation of vectors was used (Pearson or Spearman). The authors conclude that the general picture provides support for the original and for the generality of Spearman's hypothesis.

Rushton, Skuy, and Fridjohn (2003) administrated the Raven's Advanced Progressive Matrices to 294 17–23-year-old engineering students in South Africa. The sample consisted of 187 Blacks, 40 East Indians and 67 Whites. The IQ equivalent of the obtained test scores (using the 1993 US norm for 18–22 year olds) was 103, 106 and 117. Using either the point-biserial or biserial item-total correlation, the item's *g* loading was estimated. The vector of *g* loadings was then correlated, using either Pearson and Spearman correlations, with the vector of the racial differences in pass ratios in order to test Spearman's hypothesis. The results showed that when using the *g* loadings obtained from the White sample the correlations were significant (0.34–0.39) for the W–B, but not the W–I or the B–I. When using the *g* loading obtained from the minority sample, the W–B showed evidence in favor of Spearman's hypothesis (0.37–0.64). As for W–I, only the point-biserial correlation reached significance (0.29–0.32), and none of the B–I correlations reached significance. The results support the original Spearman's hypothesis but provide only limited support for the generality for other White–minority difference and none for difference between minority groups.

Helm-Lorenz, van de Vijver, and Poortinga (2003) administrated either six or four cognitive ability tests (either the SON-R or the RAKIT) and two computerized elementary cognitive tests (TAART) to 747 White majority Dutch children and 474 Dutch minority children (primarily from Turkey and Morocco), all aged 6–12. The majority mean test scores were 0.7 S.D. and 1.1 S.D., respectively (measured in majority S.D. units) above the minority group means. A joint factor analysis of the ECT measures and either one of the two cognitive ability testbatteries was used in order to estimate the *g* loadings. This vector of *g* loadings was then correlated with the vector of differences in test scores in order to test Spearman's hypothesis. The result showed correlations ranging from −0.30 to −0.36 ($p<0.01$) and −0.37 to −0.45 (n.s.) for minority and majority *g* loadings, respectively, depending on whether each 1-year age span was used or collapsed and averaged. These results are contrary to expectation, as there was less of a difference on highly *g* loaded tests. The authors further tested whether the cultural and verbal loadedness of the subtests (assessed by rating scales and the complexity of the material and instructions) was related to the difference in test scores, and found that this certainly was the case ($r=0.67$, $p<0.01$). This study does not support the Spearman's hypothesis in any form but rather suggests that culturally loaded subtests produce the difference across racial groups.

te Nijenhuis, Tolboom, Resing, and Bleichrodt (2004) administrated the RAKIT (12 tests) to a sample of children aged 7 1/2—almost 8, and gathered teacher ratings on five school topics. The sample consisted of 196 White Dutch children, and Dutch immigrants' children from the Netherlands Antilles/Surinam ($N=61$), Turkey ($N=71$) and Morocco ($N=61$). The mean ability differences, as measured by deviations from the White mean in White S.D. units, were −0.81; −1.43; and −1.83, respectively, in favor of the majority White children. A test of Spearman's hypothesis in general provided no significant support, although the findings indicated a tendency towards support for the hypothesis for the RAKIT test scores as well as for teacher ratings. The study also presented two further samples aged 5½–6 and 9½–10, of similar size and demographic composition that provided similar results.

te Nijenhuis and van der Flier (2004) tested young blue collar job applicants (mean age ranging from ca. 23 to 30, depending on subgroup) with a battery of *g* loaded safety tests. The tests measure selective attention, attentional speed, perceptual-motor ability, sensory-motor coordination ability, and precision of reactions. The sample consisted of 584 majority White Dutch, and two Dutch immigrant groups composed of 466 Surinamese/Antillean and 320 African/Turkish minority. The mean difference in ability, as measured by deviations from the White mean in White S.D. units, were −1.10 and −1.59, respectively. The authors tested Spearman's hypothesis by using the safety tests correlation with the GATB as measure of *g* loading and correlated this with the vector of racial difference on the safety tests. The results provided support for the hypothesis in terms of finding large and significant Pearson correlations of 0.67 and 0.77 ($p<0.05$).

Our review of the literature on the cross-racial generality of Spearman's hypothesis thus suggests:

1. A strong relationship between subtest *g* loadings and White–Mexican IQ differences.
2. A zero to moderate relationship between subtest *g* loadings and Whites–Asians (Japanese/Chinese) IQ differences.
3. A zero to moderate relationship between *g* loadings and White–Indians IQ differences.

4. A moderate to strong relationship between subtest *g* loadings and the IQ differences for Whites and four immigrant groups (although sometimes finding the opposite pattern).

In fairness, it is worth keeping in mind that some of these conclusions are based on only a few studies.

The purpose of the present study is to further test the cross-racial generality of Spearman's hypothesis. More specifically, we asked whether IQ differences among Americans of European origin (W) and Americans of Hispanic/Latino (H) descent are attributable to differences in *g*, while drawing upon data from two very large and different samples.

Our review of the scientific literature leads us to believe that this would be the case.

2. Sample 1

2.1. Subjects

Data for our first analysis was originally collected by the Centers for Disease Control (1988, 1989), in order to assess possible long-term effect of toxic exposure and other risks from military service in 4462 male soldiers, inducted in 1965–1971 and re-tested in 1985–1986. About half served in Vietnam and the remaining outside Vietnam, but no major differences in test results were observed. The Vietnam Experience Study (VES) sample is fairly representative of the US population with respect to education, income, occupation, and race, but subjects scoring below the 10th percentile in the pre-induction cognitive aptitude test were excluded, in accordance with a US Congress mandate. This obviously truncates the lower-end tail of the ability distribution. The age of the subjects at the second testing was 38.35 (S.D. 2.52). Factor analysis left us with valid data for 3556 W and 181 H. The Hispanic N is admittedly relatively small, but still appropriate for factor analysis, even if the outcome has to be interpreted conservatively.

2.2. Psychometric variables

The VES study operates with no less than 19 independent cognitive ability tests, all of a highly diverse nature: (Grooved Pegboard Test*, left and right hand*; Paced Auditory Serial Addition Test; Rey-Osterrieth Complex Figure Drawing, *direct copy, immediate recall and delayed recall*; Wechsler Adult Intelligence Scale-Revised, *general information and block design*; Word List Generation Test; Wisconsin Card Sort Test; Wide Range Achievement Test; California Verbal Learning Test; Army Classification Battery, *verbal and arithmetic reasoning*, *administered twice*; Pattern Analysis Test; General Information Test; Armed Forces Qualification Test). The considerable number and diversity of the tests allows for the derivation of a high-quality measure of general intelligence *g*. Five of the tests were administered at time of induction; all the remaining were administered at the second testing 17.90 (S.D.= 1.86) years after induction, on average. A more detailed description of the tests is provided elsewhere (Nyborg & Jensen, 2000). Since IQ and *g* scores are highly stable over time (Brody, 1992), we decided to include all the cognitive tests in the battery for the extraction of *g*.

2.3. Sample 2

2.3.1. Subjects

A total of 11,625 subjects (5808 males and 5817 females) aged 15–24 years (mean=19.6; S.D.=2.26) completed the American Services Vocational Aptitude Battery (ASVAB) as part of the 1979 National Longitudinal Study of Youth (NLSY1979; www.bls.gov/nls/home.htm) under standard test conditions. Factor analysis left us with valid data for 6947 W (3478 males and 3469 females) and 1704 H (836 males and 868 females).

2.3.2. Psychometric variables

The ASVAB is a 10 subtest multiple-choice test, including a wide range of both ability and knowledge tests (General Science; Arithmetic Reasoning; Word Knowledge; Paragraphs Comprehension; Numerical Operations; Coding Speed; Automobile and Shop Information; Mathematics Knowledge; Mechanical Comprehension; Electronics Information). The test has been described in details elsewhere (e.g. Evans, 1999; Legree, Pifer, & Grafton, 1996).

2.4. Statistical methods

General intelligence *g* was determined in both studies for each sample as the first un-rotated principal factor (PAF1) of two extracted factors in the total samples, thus including other racial groups besides W and H, in order to obtain the best possible *g* estimate.

We then calculated the effect measure for each test in both studies in order to express the subtest race differences in terms of d effects. This is done by subtracting W means by H means and dividing the

difference by the average S.D. for the two groups, using the formula:

$$d = (X_1 - X_2)/\sqrt{(N_1^* s_1^2 + N_2^* s_2^2)/(N_1 + N_2)}$$

where X_1 and X_2 are the means for the two groups, $N_1 + N_2$ the number of subjects in the two groups, s_1 and s_2 the standard deviation in each group, and the d-differences are set on the same scale and thus comparable (Jensen, 1998, p. 403).

We, finally, applied the method of correlated vectors to see whether the size of the subtest racial d-effect differences correlate with the g-loads of the subtests (the Jensen effect), in order to test Spearman's hypothesis. Details of the method are found in Jensen (1998, Chapter 11 and Appendix A).

3. Results of PAF analyses

The results are presented in Tables 2 and 3. The last rows indicate that W scored 0.78–0.86 S.D. above H.

The g (PAF1) loadings of the subtests were found through a Principal Axis Factor analysis (PAF) of the two subgroups within each sample. The Screeplot indicated that it was possible to extract 3 factors with eigenvalues above 1 in sample 2, and two factors in the three other samples. As the loadings of the extracted factors may vary as a function of the number of extracted factors when applying a Principal Axis Factoring, and as we wanted uniformity across samples and groups, we decided to extract two factors for all samples. After unearthing excellent factor congruence (all congruence coefficients >0.98), we averaged the loadings of the PAF1 and PAF2 across subgroups, for each sample independently, using the formula:

$$a_{\text{average}} = \sqrt{(a_1^2 + a_2^2)/2}$$

where a_1 and a_2 are the factor loadings on the specific factor for the two groups (Jensen, 1998, p. 406).

The results of the PAF for the two groups within the two samples are presented in Tables 4 and 5, as are the communalities of the subtests with two factors extracted (PAF1 and PAF2). Communalities are used here as a lower bound measure of reliability correction to reduce the likelihood that the correlation is an artifact of unreliability (see Nyborg & Jensen, 2000). The use of communalities tends to work against a confirmation of Spearman's hypothesis (see Nyborg & Jensen, 2000) but is better than no control whatsoever, and none of the databases contained measures of reliability for the single subtests.

Table 2
Sample 1: Mean differences, S.D. and effect sizes of the ability difference across Whites (N=3556) and Hispanics (N=181)

Age, tests, and PAF1	Mean		t-value	S.D.		F-ratio	d
	White	Hispanic		White	Hispanic		
Age	38.37	38.32	0.30	2.49	2.57	1.06	0.02
GTP(RH)	−73.14	−72.45	−0.79	11.41	11.43	1.00	−0.06
GTP(LH)	−76.54	−75.49	−1.09	12.69	11.00	1.33 *	−0.08
PASAT	114.15	96.41	4.75 **	48.95	50.96	1.08	0.36
CFD (copy)	33.03	32.12	3.91 **	3.00	3.74	1.55 *	0.30
CFD (immediate recall)	20.74	18.20	5.09 **	6.55	6.59	1.01	0.39
CFD (delayed)	20.84	18.47	5.05 **	6.18	6.08	1.03	0.39
WAIS-R (general information)	10.42	8.86	7.42 **	2.76	2.70	1.05	0.57
WAIS-R (block design)	10.89	9.73	6.06 **	2.53	2.21	1.32 *	0.46
WGLT	35.57	31.87	4.48 **	10.93	9.53	1.32 *	0.34
WCST	0.80	0.77	3.09 **	0.17	0.18	1.17	0.24
WRAT	62.65	60.03	2.39 *	14.34	14.18	1.02	0.18
CVLT	47.05	42.41	7.09 **	8.53	9.48	1.23 *	0.54
ACB Verbal (I)	110.68	93.97	10.42 **	21.06	20.36	1.07	0.79
ACB Verbal	120.21	103.78	10.17 **	21.07	23.61	1.26 *	0.78
ACB Arithmetic (I)	108.00	93.55	9.12 **	20.83	20.11	1.07	0.69
ACB Arithmetic	108.70	92.14	9.54 **	22.83	21.94	1.08	0.73
PAT (I)	107.11	98.12	5.37 **	22.08	20.01	1.22	0.41
GIT (I)	105.42	88.14	13.37 **	16.81	19.75	1.38 *	1.02
AFGT (I)	58.40	40.66	9.48 **	24.72	21.66	1.30 *	0.72
PAF1	0.17	−0.56	10.30 **	0.91	0.83	1.21	0.78

* p>0.05 (two-tailed).
** p>0.005 (two-tailed).

Table 3
Sample 2: Mean differences, S.D. and effect sizes of the ability difference across Whites (N=6947) and Hispanics (N=1704)

Age, tests, and PAF1	Mean		t-value	S.D.		F-ratio	d
	White	Hispanics		White	Hispanics		
Age	19.75	19.26	7.93 **	2.27	2.22	1.04	0.22
General Science	16.33	12.63	29.23 **	4.64	4.83	1.08 *	0.79
Arithmetic Reasoning	18.44	13.64	25.79 **	7.06	6.17	1.31 **	0.70
Word Knowledge	26.77	21.25	28.13 **	7.11	7.83	1.21 **	0.76
Paragraphs Comprehension	11.19	8.99	25.02 **	3.18	3.57	1.26 **	0.68
Numerical Operations	35.04	29.61	19.20 **	10.32	10.96	1.13 *	0.52
Coding Speed	47.00	40.22	16.30 **	15.36	15.57	1.03	0.44
Automobile and Shop Information	14.68	11.10	25.39 **	5.23	5.14	1.04	0.69
Mathematics Knowledge	13.81	10.32	21.12 **	6.23	5.58	1.25 **	0.57
Mechanical Comprehension	14.58	11.18	25.10 **	5.07	4.74	1.15 **	0.68
Electronics Information	11.83	8.69	28.93 **	4.025	3.98	1.02	0.78
AFQT	51.79	31.30	27.98 **	27.66	24.58	1.27 *	0.76
PAF1 (all)	0.39	−0.36	31.63 **	0.87	0.87	1.01	0.86

* $p<0.05$ (two-tailed).
** $p<0.0005$ (two-tailed).

The averaged factor loadings (PAF1 and PAF2) across the two subgroups within each of the two samples were correlated with the differences in subtest scores measured in S.D. or d units, as is common when applying the method of correlated vectors. Three types of correlations were calculated: Pearson's r, Spearman's Rho (for non-normal distribution or no interval scale), and a first-order Pearson correlation, partialling out communality (to see if the reliability of the measures affects the correlation).

4. Results of correlated vector analyses

Table 6 provides the outcome of the analysis of correlated vectors.

The Pearson and Spearman correlations between PAF1 and d were large (about 0.8) and significant in both samples. This supports the cross-racial generality of Spearman's hypothesis. However, partialling out the communalities lowered the correlation for both samples and rendered the sample 2 correlation non-significant.

Table 4
Sample 1: Factor solution

Test and eigenvalues	White		Hispanic		Communalities
	PAF1 (N=3556)	PAF2 (N=3556)	PAF1 (N=181)	PAF2 (N=181)	PAF1+PAF2 *all* subjects (N=4321)
GTP(RH)	0.31	0.16	0.18	0.21	0.13
GTP(LH)	0.32	0.17	0.14	0.17	0.14
PASAT	0.53	−0.04	0.49	0.03	0.32
CFD (copy)	0.45	0.31	0.27	0.36	0.32
CFD (immediate recall)	0.56	0.68	0.55	0.66	0.78
CFD (delayed)	0.56	0.68	0.53	0.69	0.79
WAIS-R (general information)	0.75	−0.21	0.74	−0.13	0.62
WAIS-R (block design)	0.63	0.25	0.48	0.24	0.50
WGLT	0.52	−0.12	0.50	−0.18	0.26
WCST	0.41	0.06	0.33	−0.03	0.21
WRAT	0.74	−0.30	0.59	−0.23	0.63
CVLT	0.47	0.06	0.55	−0.01	0.24
ACB Verbal (I)	0.81	−0.33	0.78	−0.31	0.78
ACB Verbal	0.81	−0.30	0.82	−0.30	0.77
ACB Arithmetic (I)	0.79	−0.18	0.74	−0.16	0.68
ACB Arithmetic	0.79	−0.11	0.78	−0.02	0.67
PAT (I)	0.69	0.15	0.54	0.14	0.51
GIT (I)	0.65	−0.17	0.60	−0.14	0.49
AFGT (I)	0.84	−0.02	0.70	−0.06	0.72
Eigenvalues	7.66	1.58	6.31	1.52	

Table 5
Sample 2: Factor solution

Tests and eigenvalues	Hispanics		White		Communalities
	PAF1 (N=1704)	PAF2 (N=1704)	PAF1 (N=6947)	PAF1 (N=6947)	(PAF1+PAF2) all subjects (N=11,625; White, Black, Hispanics)
General Science	0.85	0.12	0.85	0.14	0.78
Arithmetic Reasoning	0.83	−0.04	0.85	−0.09	0.74
Word Knowledge	0.86	−0.05	0.85	−0.08	0.78
Paragraphs Comprehension	0.80	−0.17	0.77	−0.21	0.70
Numerical Operations	0.65	−0.47	0.66	−0.46	0.69
Coding Speed	0.58	−0.45	0.57	−0.46	0.59
Automobile and Shop Information	0.70	0.39	0.63	0.53	0.71
Mathematics Knowledge	0.78	−0.09	0.79	−0.19	0.66
Mechanical Comprehension	0.75	0.30	0.75	0.36	0.73
Electronics Information	0.80	0.31	0.79	0.37	0.78
Eigenvalues	5.84	0.82	5.73	1.09	

Spearman's hypothesis thus stands the critical test only in sample 1.

5. Discussion

The observed general intelligence *g* difference between W and H of less than 1 S.D. accords well with our literature review and with the observation by Sandoval (1982).

The present study further provides support for the generality of Spearman's hypothesis with respect to race differences in *g* between W and H. However, the apparent strong support found in both samples (r about 0.8; $p<0.05$) is curbed by the fact that the hypothesis could be defended only using sample 1, after partialling out communalities. In other words, the results from sample 1 dovetail nicely with our literature review of the few relevant studies on this topic, but the analysis of sample 2 failed to document the expected significant relationship between *g* loadings and ability differences for W and H.

The reasons for sample discrepancies and incomplete confirmation of the Spearman's hypothesis may be attributed to a number of factors.

Table 6
Samples 1 and 2: *d* vectors correlated with PAF1 and PAF2

Sample 1	Average PAF1 and *d*	Average PAF2 and *d*
19 subtests		
Pearson's r	0.82 ($p<0.0005$)	−0.08 ($p<0.1$)
Spearman's Rho	0.82 ($p<0.0001$)	−0.10 ($p<0.1$)
Pearson's with communalities partialled out	0.69 ($p<0.005$)	−0.58 ($p<0.025$)
Sample 2	**Average PAF1 and *d***	**Average PAF2 and *d***
10 subtests		
Pearson's r	0.81 ($p<0.005$)	−0.52 ($p<0.1$)
Spearman's Rho	0.83 ($p<0.005$)	−0.48 ($p<0.1$)
Pearson's r with communalities partialled out	0.30 ($p<0.1$)	−0.25 ($p<0.1$)

All p-values are two-tailed.

5.1. Sample size

Sample 2 contains a much larger number of Hs than sample 1. This means that the results from this sample should weigh more heavily than the outcome of the sample 1 analysis, due to potentially fewer measurement errors. However, given this, then the inconsistencies across samples must be attributed to measurement error, and the results are then inconsistent with the literature on the subject, unless it is assumed that the findings by Sandoval (1982; $N\approx300$) are also invalid due to "too few" subjects. Then again, even if $N=150$ is a small sample for factor analysis, 300 subjects should suffice to provide valid results. Furthermore, it is highly unlikely that the measurement error introduced in the factor analysis due to lack of subjects should be identical across two studies. We accordingly suspect that mere variation in sample size cannot explain the diverging results.

5.2. Number and nature of the variables

Sample 1 was subjected to more subtests than sample 2. Could this explain the difference in results? Not likely. Both samples have a sufficient number of sufficiently diverse tests for factor analysis. Adding further tests beyond these lower limits would not

seriously affect the analysis. One could argue that the nature of the test used in sample 2 is more verbally and culturally loaded (Roberts et al., 2000), and thus produces less support for the generality of Spearman's hypothesis (e.g. Helms-Lorenz, Van de Vijver, & Poortinga, 2003). However, if this were the case, then a difference across samples would be apparent before the partialling of reliability, which was not the case.

5.3. Age of the subjects

Age cannot explain the differences in results. Thus, the Sandoval (1982) sample was composed of children aged 5–11, whereas the age ranged between 20 and 41 years in the present study. If age is a factor, then the generality of Spearman's hypothesis is confirmed for children and middle-aged adults, but not for teenagers/young adults. This seems highly unlikely.

5.4. Restriction of range

Given that the *g* loadings of the W sample subtests were uniformly higher than for H, one could attribute the diverging results to restriction of range in sample 2. However, sample 2 results actually confirmed the hypothesis *before* partialling out communalities. If restriction of range were an issue, one would expect the hypothesis to be disconfirmed before this partialling out.

5.5. Strictness of the procedure

As pointed out by Nyborg and Jensen (2000), the use of communalities as a substitute for reliability is a very stringent procedure that tends to work against the hypothesis. The question therefore arises whether the initial confirmation is more correct than the subsequent disconfirmation after correcting for reliability. Nyborg and Jensen (2000) argued that the correction made by communalities is preferable to no correction at all, and we therefore suspect that although the corrected correlations are probably underestimating the actual size of the correlation, it probably is a more valid approach than just presenting the initial correlations without any correction. Furthermore, one would expect the stringent procedure of using communalities as an estimate of reliability to work equally against the hypothesis for both samples and not just for one. However, when looking at the ratio between the variance of the first and second extracted factors in the two samples, it becomes apparent that it is larger in the second sample. This suggests, in turn, that the communalities are to a large extent composed of variance from the first factor. Partialling out the communalities would then remove more variance from the first extracted factor in the second sample than in the first sample, thereby providing an explanation for the lack of confirmation after correction. This is, in our opinion, the most likely candidate for explaining the differences across samples.

6. Concluding remarks

The present study provided support for the cross-racial generality of Spearman's hypothesis by showing that, for sample 1, the differences in general intelligence *g* across Whites and Hispanics are highly attributable to differences in the *g* factor. However, results from sample 2 were less clear-cut: findings support the generality before correction for reliability, but not after correction. After discussing a number of possible confounders in sample 2, we came to the conclusion that the strict procedure of using communalities rather than reliability coefficients, probably explains the diverging results across samples. Obviously, more experimental research is needed to clarify the W–H generality of the hypothesis.

The present study cannot spread light on the nature-nurture question of the observed subtest score differences across the W and H groups. Whereas differences in IQ may be partly explained by educational factors (Ceci, 1991), test knowledge/practice (Jensen, 1980), or differences in lower stratum cognitive factors (Jensen, 1998; Nyborg & Jensen, 2000), the present study suggests that the observed IQ differences across the W and H groups are primarily attributable to differences in *g*. This raises a further question: Where do race differences in *g* come from?

According to Jensen's default hypothesis (Jensen, 1998), differences in *g* between races are caused by the same agents that explain individual differences in *g* within races. Since genetic differences among individuals account for about half the variance in IQ scores (Plomin, DeFries, McClearn, & Rutter, 1997) and possibly more for adults, then race differences in *g* would by analogy be no less due to genetic differences and no more due to environmental factors. However, Brody (2003) argues that race differences have a larger environmental component than commonly assumed, namely, that at least some of the differences can be "remedied" by improving the environment of minority group, and that, in general, the evidence for Jensen's default hypothesis is less than convincing. Then again,

the notion that race differences in *g* are simply a function of aggregated individual differences in *g* has long received broad and strong psychometric, behavior-genetic, and evolutionary support (Rushton, 2003). Rushton and Jensen (2005a) recently published a review of research on race differences in cognitive ability, indicating that the 15–18 point average IQ difference between Blacks and Whites (about 1.1 standard deviation) has not narrowed down since they were first measured nearly 100 years ago and, most importantly, has not changed in accordance with the concomitant large changes in Black standards of living and level of education. They find that the " … IQ differences are attributable to differences in brain size more than to racism, stereotype threat, item selection on tests …" (Rushton & Jensen, 2005b, p. 328) and all other suggestions offered by the critiques. Large-scale intervention programs such as the well-known "Head Start" project have not produced lasting changes, and the evidence for the predominantly biological and genetic nature of *g* is growing by the day (Jensen, 1998). The findings by Rushton and Jensen is supported by Gottfredson (2005) who found the discussion of the W–B IQ gap (as presented by Rushton & Jensen, 2005a) consistent with a hereditarian 50% genetic causation theory." However, this view advocated by Jensen and Rushton is not without it critics. Sternberg (2005) comments directly on the Rushton and Jensen (2005a) paper and sees no public-policy implications arise from their analysis of the alleged genetic bases for average race differences in IQ. Nisbett (2005) claims that the W–B IQ gap has, in fact, narrowed in recent years and questions a hereditarian interpretation of the gap. Suzuki and Aronson (2005) emphasize traditional cultural and environmental factors in order to explain race differences in IQ. Rushton and Jensen (2005b) counter these critical voices one by one, and finally suggest: "It is time to stop committing the "moralistic fallacy" that good science must conform to approved outcomes" (Rushton & Jensen, 2005b, p. 328).

The present study does not allow us to unequivocally conclude that the observed cognitive W–H difference is attributable to differences in general intelligence, *g*, nor whether it accords with Jensen's default hypothesis. However, it seems reasonable to speculate that W–H differences can be explained by the same factors that explain W–B differences and that explain individual differences in *g*. Whereas evidence for the default hypothesis for the W–B difference continues to amass, we are only at the beginning of actually testing whether it generalizes to other races as well.

References

Ashton, M., & Lee, K. (2005). Problems with the method of correlated vectors. *Intelligence*, *33*, 431–444.

Brody, N. (1992). *Intelligence* (2nd ed.). San Diego, CA: Academic Press.

Brody, N. (2003). Jensen's genetic interpretation of racial differences in intelligence: Critical evaluation. In H. Nyborg (Ed.), *The scientific study of general intelligence: Tribute to Arthur R. Jensen* (pp. 397–410). New York: Pergamon.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor analytic studies.* New York: Cambridge University Press.

Ceci, S. J. (1991). How much does schooling influence intellectual development and its cognitive components. *Developmental Psychology*, *27*, 703–722.

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

Centers for Disease Control (1989). *Health status of Vietnam veterans: Vol IV. Psychological and neuropsychological evaluation.* Atlanta, GA: Center for Environmental Health and Injury Control.

Evans, M. G. (1999). On the asymmetry of *g*. *Psychological Reports*, *85*, 1059–1069.

Gottfredson, L. S. (2005). What if the hereditarian hypothesis is true? *Psychology, Public Policy, and Law*, *11*, 311–319.

Helms-Lorenz, M., Van de Vijver, F. J. R., & Poortinga, Y. H. (2003). Cross-cultural differences in cognitive performance and Spearman's hypothesis: g or c? *Intelligence*, *31*, 9–29.

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

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

Legree, P. J., Pifer, M. E., & Grafton, F. C. (1996). Correlations among cognitive abilities are lower for higher ability groups. *Intelligence*, *23*, 45–57.

Lynn, R., & Owen, K. (1994). Spearman's hypothesis and test score differences between Whites, Indians and Blacks in South Africa. *The Journal of General Psychology*, *121*, 27–36.

Nagoshi, C. T., Johnson, R. C., DeFries, J. C., Wilson, J. R., & Vandenberg, S. G. (1984). Group differences and first principal-component loadings in the Hawaii family study of cognition: A test of the generality of Spearman's hypothesis. *Personality and Individual Differences*, *5*, 751–753.

Nisbett, R. E. (2005). Heredity, environment, and race difference in IQ: A commentary on Rushton and Jensen. *Psychology, Public Policy, and Law*, *11*, 302–310.

Nyborg, H., & Jensen, A. R. (2000). Black–White differences on various psychometric tests: Spearman's hypothesis tested on American Armed Services Veterans. *Personality and Individual Differences*, *28*, 593–599.

Owen, K. (1992). The suitability of Raven's Standard Matrices for various groups in South Africa. *Personality and Individual Differences*, *13*, 149–159.

Plomin, R., DeFries, J. C., McClearn, G. E., & Rutter, M. (1997). *Behavioral genetics* (3rd ed.). New York: Freemann.

Roberts, R. D., Goff, G. N., Anjoul, F., Kyllonen, P. C., Pallier, G., & Stankov, L. (2000). The Armed Services Vocational Aptitude Battery (ASVAB). Little more than acculturated learning (Gc)!? *Learning and Individual Differences*, *12*, 81–103.

Rushton, J. P. (2002). Jensen effect and African/Coloured/Indian/White differences on Raven's Standard Progressive Matrices in South Africa. *Personality and Individual Differences*, *33*, 1279–1284.

Rushton, J. P. (2003). Race differences in g and the "Jensen effect". In H. Nyborg (Ed.), *The scientific study of general intelligence: Tribute to Arthur R. Jensen* (pp. 147–186). New York: Pergamon.

Rushton, J. P., & Jensen, A. R. (2005a). Thirty years of research on race differences in cognitive ability. *Psychology, Public Policy, and Law*, *11*, 235–294.

Rushton, J. P., & Jensen, A. R. (2005b). Wanted: More race realism, less moralistic fallacy. *Psychology, Public Policy, and Law*, *11*, 328–336.

Rushton, J. P., Skuy, M., & Fridjohn, P. (2002). Jensen effects among African, Indian, and White engineering students in South Africa on Raven's standard Progressive Matrices. *Intelligence*, *30*, 409–423.

Rushton, J. P., Skuy, M., & Fridjohn, P. (2003). Performance on Raven's Advanced Progressive Matrices by African, East Indians, and White engineering students in South Africa. *Intelligence*, *31*, 123–137.

Sandoval, J. (1982). The WISC-R factorial validity for minority groups and Spearman's hypothesis. *Journal of School Psychology*, *20*, 198–204.

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

Sternberg, R. J. (2005). There are no public-policy implications: A reply to Rushton and Jensen. *Psychology, Public Policy, and Law*, *11*, 295–301.

Suzuki, L., & Aronson, J. (2005). The cultural malleability of intelligence and its impact on the racial/ethnic hierarchy. *Psychology, Public Policy, and Law*, *11*, 320–327.

te Nijenhuis, J. (1997). *Comparability of test scores for immigrants and majority group members in the Netherlands*. Unpublished doctoral dissertation. Amsterdam, The Netherlands: Vrije Universteit.

te Nijenhuis, J., Evers, A., & Mur, J. P. (2000). Validity of the differential aptitude test for the assessment of immigrant children. *Educational Psychology*, *20*, 99–115.

te Nijenhuis, J., Tolboom, E. A. M., Resing, W. C. M., & Bleichrodt, N. (2004). Does cultural background influence the intellectual performance of children from immigrant groups. Validity of the RAKIT intelligence test for immigrant children? *European Journal of Psychological Assessment*, *20*, 10–26.

te Nijenhuis, J., & van der Flier, H. (1997). Comparability of GATB scores for immigrants and majority group members: Some Dutch findings. *Journal of Applied Psychology*, *82*, 675–687.

te Nijenhuis, J., & van der Flier, H. (2004). The use of safety suitability tests for the assessment of immigrants and majority group job applicants. *International Journal of Selection and Assessment*, *12*, 230–242.