Journal of
Intelligence

MDPI

Article

Spearman's Hypothesis Tested on Black Adults: A Meta-Analysis

Jan te Nijenhuis * and Michael van den Hoek

Work and Organizational Psychology, University of Amsterdam, Weesperplein 4, 1018 XA Amsterdam, The Netherlands; michael.vandenhoek@student.uva.nl
* Correspondence: JanteNijenhuis@planet.nl; Tel.: +31-20-5256860

Academic Editor: Paul De Boeck
Received: 29 December 2015; Accepted: 27 May 2016; Published: 1 June 2016

Abstract: Blacks generally score significantly lower on intelligence tests than Whites. Spearman's hypothesis predicts that there will be large Black/White differences on subtests of high cognitive complexity, and smaller Black/White differences on subtests of lower cognitive complexity. Spearman's hypothesis tested on samples of Blacks and Whites has consistently been confirmed in many studies on children and adolescents, but there are many fewer studies on adults. We carried out a meta-analysis where we collected the existing tests of Spearman's hypothesis on adults and collected additional datasets on Black and White adults that could be used to test Spearman's hypothesis. Our meta-analytical search resulted in a total of 10 studies with a total of 15 datapoints, with participants numbering 251,085 Whites and 22,326 Blacks in total. For all these data points, the correlation between the loadings of a general factor that is manifested in individual differences on all mental tests, regardless of content (*g*) and standardized group differences was computed. The analysis of all 15 data points yields a mean vector correlation of 0.57. Spearman's hypothesis is confirmed comparing Black and White adults. The differences between Black and White adults are strongly in line with those previously found for children and adults; however, because of lack of access to the original data, we could not test for measurement invariance.

Keywords: Spearman's hypothesis; *g*; intelligence; meta-analysis

1. Introduction

IQ test scores are well-known as excellent predictors of many economic, educational, and social criteria [1], and therefore group differences in mean intelligence are of great interest. By far the most extensively researched is that between the two largest populations in the United States: Whites, or Caucasians, and Blacks, or African Americans [1]. On average, the American Black population scores below the White population by about 1.2 standard deviations (*SD*s), or 18 IQ points. Large-scale research on cultural bias in the test instruments against Blacks simply has not shown convincing proof [2].

Te Nijenhuis, Al-Shahomee, van den Hoek, Allik, Grigoriev, and Dragt [3] describe how a well-established empirical finding—the manifold of positive correlations among measures of various mental abilities—is generally considered to be evidence of a general factor in all of the measured abilities. The use of the method of factor analysis makes it possible to determine the degree to which each of the variables is correlated with the factor that is common to all the variables in the analysis. This was termed *g* by Spearman and meant to represent a general factor that is manifested in individual differences on all mental tests, regardless of content [1]. Spearman's *g* is usually defined operationally as the loading on the first unrotated factor in a principal-axis factor analysis of a varied set of IQ tests [4]. Tests with high *g* loadings demand higher cognitive complexity, and tests with low *g* loadings demand lower cognitive complexity [3,5].

Jensen [1] devised the Method of Correlated Vectors (MCV) to empirically test which phenomena are linked to the *g* factor. It involves calculating two vectors and then correlating them with each other. The first vector consists of the correlations of subtests of an IQ battery with the general factor of intelligence, their *g* loadings. The second vector consists of the relation of each of those same subtests with the variable in question; it could be a correlation of a phenomenon with the various subtests of the IQ battery (*r*), or it could be the difference between two groups on all the subtests of the IQ battery (*d*).

Spearman [6] observed that there are large Black/White differences on some subtests of an IQ battery, yet on other subtests there are much smaller Black/White differences. He suggested that these differences might be a function of each test's *g* loading, with large differences on subtests with a high *g* loading and smaller differences on subtests with low *g* loadings. This hypothesis is now known as Spearman's hypothesis. Jensen [1] distinguished between the "strong form of Spearman's hypothesis" and the "weak form of Spearman's hypothesis". The first says that the mean Black/White IQ differences are *solely* due to differences in the hypothesized realistic *g*; the second says that the mean Black/White IQ differences are *mainly* due to differences in the hypothesized realistic *g* (implying there are additional sources of Black/White IQ differences present). Jensen chose to test the weak form of Spearman's hypothesis, and this weak form has been confirmed in many studies, which means that the Black/White differences are differences in *g* to a strong degree [1].

Te Nijenhuis *et al.* describe how Spearman's hypothesis has also been studied using methods other than intelligence tests. First, there are elementary cognitive tasks (ECTs), which measure the time it takes a person to process information presented in very simple tasks. The chronometric variables derived from such ECTs show clear Black/White differences that are predicted by their *g* loadings [7]. Second, Spearman's hypothesis has also been studied using Situational Judgment Tests (SJTs) and Assessment Center (AC) exercises, which are widely used in selection for organizations. Whetzel, McDaniel, and Nguyen's [8] meta-analysis shows that group differences in SJT performance are largely explained by the cognitive loading of the SJT. Goldstein, Yusko, Braverman, Smith, and Chung [9] tested whether the cognitive complexity of an AC exercise was a predictor of group differences, and their findings are in line with Spearman's hypothesis. Goldstein, Yusko, and Nicolopoulos [10] concluded that clear group differences emerged for a majority of more cognitively-loaded managerial competencies, such as judgment, whereas much smaller differences were associated with the majority of the less cognitively-loaded competencies, such as human relations.

Te Nijenhuis *et al.* describe how Spearman's hypothesis has also been tested for Hispanic, Native-American, Asian-American, and Native-Hawaiian groups. Outside of the US, Spearman's hypothesis has been tested in the Netherlands, South Africa, Zimbabwe, Asia, and Serbia. In the majority of these cases Spearman's hypothesis was strongly confirmed.

Te Nijenhuis *et al.* describe how Rushton tested Spearman's hypothesis in a series of studies at the item level using the various versions of Raven's Progressive Matrices (RPM) in Africa and in Serbia [11–15]. The *g* loadings of items were operationalized as the items' correlation with the total score on the RPM, which has a strong correlation with the general factor of intelligence. The difference scores of items (*d*) were operationalized as the difference in pass rates between groups. It was generally found that group differences were greater on those items of the RPM with the highest item-total correlations, which are the best measures of general factor of intelligence, which counts as a confirmation of Spearman's hypothesis. More recently, a White Spanish sample was compared with a sample of Moroccans but did not report a clear confirmation of Spearman's hypothesis [16]. Most recently, a number of studies by te Nijenhuis and co-authors, using a large number of datasets, generally showed clear and often strong confirmations of Spearman's hypothesis at the item level [3,17–19].

An interesting recent paper by Ganzach [20] did not explicitly test Spearman's hypothesis using the Method of Correlated Vectors, but contrasted scores on Wechsler subtests Digit Span Forward (DSF) and Digit Span Backward (DSB), going back to Jensen's early work on group differences [21] where he showed that the Black/White difference was much larger on DSB than on DSF. Ganzach re-examined Jensen and Figueroa's results on the basis of a large, nationally representative database.

He replicated earlier findings by showing that the difference between Blacks and Whites is larger on DSB than on DSF. However, the results were not generalizable to Hispanics, where the difference between Whites and Hispanics was actually larger on DSF than on DSB. For a more detailed discussion of these findings, see David [22] and Ganzach [23].

Testing Spearman's hypothesis using the Method of Correlated Vectors has met with criticism [24,25]. Woodley, te Nijenhuis, Must, and Must [26] argue that most of the criticism of the MCV rests on two problematic premises. First, it has been made clear by Jensen [1] that one should use fairly representative samples, that a large enough number of tests should be used, and these tests should not all be similar—for instance, only reasoning tests—but must also be diverse in terms of content. A study by Ashton and Lee [24] shows that analyzing unbalanced collections of tests result in outcomes that make little sense, but they simply ignored the fact that Jensen explicitly warned researchers about the use of unbalanced samples. Second, Jensen [1] shows that there are four statistical artifacts that strongly attenuate the outcomes of the MCV, such as restriction of range and unreliability. This means that Jensen was well aware of fundamental weaknesses in MCV and he showed that controlling for them strongly increased the value of the resulting correlations between the *g* vector and the *d* vector. Dolan's [25] finding that small samples in some cases yield unreliable outcomes comes as no surprise [26].

Woodley *et al.* argue that MCV should be combined with psychometric meta-analysis [27] because it has several advantages. First, it allows the use of all published datasets. Second, it allows importing the best available *g* loadings from other datasets, thereby strongly reducing the unreliability. Third, there is information on the variance between studies, which is generally large. Fourth, corrections for several important statistical artifacts can be carried out. Dolan [25] advises the use of Multigroup Confirmatory Factor Analysis (MGCFA) instead of MCV, but then all the advantages listed above disappear (see [26] for a detailed description).

Jensen carried out a large number of tests of Spearman's hypothesis (see [1] for a review). Jensen [7] states that seven methodological requirements for the testing of Spearman's hypothesis have to be met:

1. The samples should not be selected on any highly *g*-loaded criteria.
2. The variables should have reliable variation in their *g* loadings.
3. The variables should measure the same latent traits in all groups. The congruence coefficient of the factor structure should have a value of >0.85.
4. The variables should measure the same *g* in the different groups; the congruence coefficient of the *g* values should be >0.95.
5. The *g* loadings of the variables should be determined separately in each group. If the congruence coefficient indicates a high degree of similarity, the *g* loadings of the different groups should be averaged.
6. To rule out the possibility that the correlation between the vector of *g* loadings (V_g) and the vector of mean differences between the groups or effect sizes (V_{ES}) is strongly influenced by the variables' differing reliability coefficients, V_g and V_{ES} should be corrected for attenuation by dividing each value by the square root of its reliability.
7. The test of Spearman's hypothesis is the Pearson correlation (*r*) between V_g and V_{ES}. To test the statistical significance of *r*, Spearman's rank order correlation (r_s) should be computed and tested for significance.

However, Jensen [1] shows many instances where *g* loadings and effects sizes (*r* or *d*) are correlated not using individual-level data, but group-level data reported in individual studies, so that it is not possible to test whether the seven methodological requirement were met. Usually these individual studies did not report *g* loadings, in which case Jensen generally used *g* loadings from other sources, such as manuals of IQ batteries based upon high-quality, representative samples. Jensen therefore carried out studies where he used an elaborate procedure [7] and studies where he used a more simplified procedure. It was not stated explicitly, but Jensen's requirements for comparability of test

scores when comparing Black and White samples in the large majority of cases were met, so it is possible that Jensen [1] saw it as less pressing to explicitly test for the comparability of test scores in new studies.

It is also possible that when using a simplified procedure Jensen traded quantity for quality: the datasets were less thoroughly analyzed, but it was much easier now to add the outcomes of the analyses on all kinds of new datasets to the literature, which could increase the chances of advancing scientific discussions. In case one wants to carry out an analysis on studies for which the individual-level data are unavailable and combine all these studies in a meta-analysis, the simplified procedure reported in Jensen [1] has to be applied. This combination of a simplified procedure and a meta-analysis has already been successfully applied in several studies, some of them often cited [28–35].

In the simplified procedure, each subtest is not corrected for unreliability, but meta-analytical corrections for unreliability are applied to the vectors [36]. Additionally, significances are not computed for each individual dataset, as the combined datasets become so large that the mean correlation will always or virtually always become significant [36]. In these studies a strong increase in the number of studies that can be analyzed is traded off for a less thorough testing procedure. An important advantage is that all the studies can now be combined into a meta-analysis, so powerful meta-analytical techniques can be applied, allowing the drawing of strong conclusions.

Dolan [25] is of the opinion that Multigroup Confirmatory Factor Analysis is preferable to MCV when testing Spearman's hypothesis, because it allows for a strong test of measurement invariance. To satisfy the critics of MCV, ideally both MCV and MGCFA should be applied to the data. However, there is a fundamental problem in that MGCFA is so demanding of the data that, in all likelihood, only a fraction of available datasets can be analyzed using MGCFA, whereas in principle all or virtually all datasets can be analyzed with MCV. Therefore, even if all datasets are available, a comparison of the two methods cannot be made in the large majority of cases, impeding a thorough evaluation of the merits of MGCFA. Obviously, a statistical technique that can only be applied to a very small selection of datasets has strong drawbacks.

In sum, Spearman's hypothesis was confirmed in the large majority of comparisons of various groups and for all assessment instruments studied and most studies have been carried out comparing Blacks and Whites in the US. However, a careful look at those US Black/White comparisons makes it clear that, by far, most research participants in these studies are children and adolescents [1]. Before the present study, the total literature of Spearman's hypothesis tested on Black and White adults in the US consists of only six datasets on adults, and five of these six studies are reported in Jensen [37] and one of these six is reported in Nyborg and Jensen [38]; the correlations for these studies ranged from $r = 0.30$ to 0.81. Thus, most studies on Spearman's hypothesis are not representative of the normal working-age population and more studies are needed to see whether the abundant findings from children and adolescents generalize to adults.

In the present study we carried out a meta-analysis where we collected the existing tests of the weak form of Spearman's hypothesis on adults and collected additional datasets on Black and White adults in the US that could be used to test the weak form of Spearman's hypothesis. We expected to find a strong confirmation of the weak form of Spearman's hypothesis for adults, just as was already found for children and adolescents.

2. Method

2.1. Meta-Analysis

In their influential book, Hunter and Schmidt [39] plead for the use of meta-analysis, which is best described as the aggregation of data from different studies and datasets, which are then corrected for statistical and study artifacts. In this study we carry out a bare-bones meta-analysis where we correct only for sampling error in the data, the error that is introduced into data due to the usage of small samples in studies. This correction was carried out using the Hunter and Schmidt Meta-Analysis Programs [40].

2.2. Inclusion Criteria

The first requirement for a study to be included in the present meta-analysis was having at least four tests or subtests to which the Method of Correlated Vectors could be applied. Second, the samples in the studies should not be selected on a highly *g*-loaded variable (e.g., referral or gifted samples [7]). Third, the average age of the participants had to be 18 or older, which means the sample could include some participants younger than 18.

2.3. Searching and Screening Studies

Several search strategies were used. For the digital search, the following electronic databases were used: Google Scholar, ProQuest, PsycINFO, and CataloguePlus (Primo). The keywords and phrases used to find further data for the data of Black adults were: "Black adult(s)", "Black and White adults", "Negro adults", "Spearman's hypothesis adults", "Spearman's hypothesis Black", "Spearman's hypothesis Negro", and "Spearman's hypothesis workforce". These keywords were combined with the following keywords: "intelligence", "IQ", "mental ability", "mental capacity", "cognitive ability", "aptitude", "competence", "differences", "WAIS" (Wechsler Adult Intelligence Scale), "KAIT" (Kaufman Adult Intelligence Test), and "Woodcock Johnson". To supplement the data we found, we searched the references of the studies already obtained to find other studies with potential data on Black adults. This search resulted in a total of 10 studies that are useable for our analysis, with a total of 15 data points.

2.4. Description of Available Data

An overview of the studies used for analysis can be seen in Table 1. Several of these studies focused on testing Spearman's hypothesis, but for four of the studies we tested Spearman's hypothesis ourselves using data reported at the group level. The most thorough test of Spearman's hypothesis requires that data are available at the individual level. However, in searching for studies for the present meta-analysis the present authors found that the individual-level data from published studies are generally impossible to get hold of, and therefore we focused on data that were reported at the group level. As noted above, Dolan [25] suggests using MGCFA to test Spearman's hypothesis, but since all data in this study are only available at the group level, it simply is not possible to use MGCFA to analyze these data. Instead, we have opted to use the Method of Correlated Vectors.

Table 1. Information on Studies Used in Current Meta-Analysis.

Original Publication	Previously Used to Test Spearman's Hypothesis	Data Availability	Sample Background
Carretta [41]	No	Group Level Only	Air Force applicants
Murray [42]	No	Group Level Only	Varied (Nationally representative samples)
Kaufman *et al.* [43]	No	Group Level Only	Varied (Stratified sample)
Sternberg [44]	No	Group Level Only	Mostly college students with some high school students
Department of Defense [45] [1]	Yes; In: Jensen [37]	Group Level Only	Varied (Representative sample)
Department of Labor [46] [1]	Yes; In: Jensen [37]	Group Level Only	Varied (33 different occupational samples)
Hennessy and Merrifield [47] [1]	Yes; In: Jensen [37]	Group Level Only	High School seniors
National Longitudinal Study [1,2]	Yes; In: Jensen [37]	Group Level Only	Varied (Stratified sample)
Nyborg and Jensen [38]	Yes	Group Level Only	Males in the Armed Forces
Veroff *et al.* [48] [1]	Yes; In: Jensen [37]	Group Level Only	Cross section of population Detroit

[1] These studies were taken from Jensen [37]; [2] Reference not given in Jensen [37].

2.5. Method of Correlated Vectors

As stated by Arthur Jensen [1], the Method of Correlated Vectors allows one to correlate the cognitive difficulty of a task with a secondary variable of interest such as ethnicity or sex. The *g* vector consists of the *g* loadings of the subtest of an IQ battery, while the second vector is often an effect size in the form of a correlation or an estimation of the difference between two groups. The Method of Correlated Vectors consists of taking the column with the *g* loading of each subtest in an intelligence battery and correlating them with the column of the effect size of the secondary variable of interest on those same subtests [1]. In cases where Spearman's hypothesis is tested, this effect size is often expressed in the *d* score, an estimate of the standardized difference between groups.

The *d* score was calculated by subtracting the score of the lower scoring group from the score of the higher scoring group. These differences in subtest scores between the groups are then correlated with the *g* loadings of the subtest. A strong positive correlation indicates that the difference between groups on the subtests becomes larger as the *g* loading of a subtest increases, a strong negative correlation indicates that the differences between groups on the subtest become smaller as the *g* loading of the subset increases, and a weak or non-existent correlation indicates there is no relation between the differences between groups and *g* loading. In this paper the Method of Correlated Vectors is indicated as $r\ (d \times g)$.

2.6. Calculating d

As is usual in testing Spearman's hypothesis, to calculate *d* the scores of the lower scoring Black group were subtracted from the scores of the higher scoring White group, and this difference was then divided by the highest-quality estimate of the standard deviation available, usually the standard deviation of a standardization sample.

2.7. Choice of SD Used in Calculating the Difference Scores (d)

The selection of standard deviations is very important for calculating the correct effect size, the standardized difference between groups, and therefore the following procedure was used. Whenever possible, the standard deviations were taken from nationally representative standardization samples or norming samples for the tests used in the study. This is the preferred option since the standard deviation of a large and representative sample is closer to the population standard deviation than a small study sample, and thus helps to give a more accurate indication of the effect size. However, for some tests no such samples could be obtained and, in that case, the standard deviations of the largest group were used to compute *d*, since a larger group would still have more reliable standard deviations than a small group. If both groups were of equal size, the *SD* from the majority group was used since it is more likely to be representative of the population *SD* than that of the minority group.

2.8. Selecting g Loading for Calculating r (d × g)

The selection of the correct *g* loading for calculating $r\ (d \times g)$ is important because the *g* loading based on a large and representative sample will be more representative of the *g* loading of the population. Whenever possible, the *g* loadings of the subtests in an intelligence battery are calculated using the data of nationally representative norming or standardization samples. However, some studies use samples that are not representative of the population and other studies use non-standard tests. In these cases, a choice has to be made as to what *g* loadings to use. While large samples are still preferable, they might not always be a good fit for the sample in the study. For example, some datasets for calculating *g* might have a large sample but might not be representative of the sample used in the study (e.g., a large difference in age), while other datasets have a small sample but are a much better representation of the sample used in the study.

Preference was given to using larger samples as long as they were relatively representative, however, if this was not possible or not appropriate for the data, a different solution was used.

In a few cases the *g* loadings were taken from *g* loadings mentioned in the study or calculated using intercorrelations that were given in the study. This was often the case for novel or less well-studied intelligence test batteries.

2.9. Correcting for Unequal Group Sizes in a Datapoint

In their influential book on psychometric meta-analysis, Hunter and Schmidt [39] use the sum of all participants in all groups from a study that is used as a datapoint in a meta-analysis as the value of the total sample size. However, in the data points in the current study there was often a large disparity between group sizes; for instance, quite often samples report data on 100 Blacks and 1000 Whites. A sample of 100 has quite substantial sampling error, whereas a sample of 1000 indicates a much smaller sampling error.

What is a good indicator of the sample size of such a datapoint combining two datasets? The strictest choice would be to simply use the value of the smallest sample. However, this would ignore the positive influence of the much larger sample on the sampling error of the datapoint. A comparison could be made with testing the means of samples of unequal size for significance: A difference between samples of 900 and 100 reaches significance less quickly than the difference between samples of 500 and 500, notwithstanding the fact that the total sample size (*N*) is equal. Stated differently, the increase in precision for the sample of 900 does not outweigh the decrease in precision for the sample of 100. A harmonic *N* takes this into account.

There are several formulas for harmonic *N* that could be used. A common formula is $\frac{N}{\frac{1}{x1}+\frac{1}{x2}+\cdots\frac{1}{xn}}$ where *N* is the number of groups and x_i is the size of each individual group [49]. The advantage of this formula is that, for a datapoint with samples of 100 and 900, the value of the harmonic $N = 180$, which is quite close to the value of the smallest sample, indicating a quite strong sampling error (see Table 2). However, the disadvantage of this formula is that for a datapoint with samples of 15 and 15, the total sample size is only 15 and that, for a datapoint with samples of 500 and 500, the total sample size is only 500 (see Table 2).

Table 2. Various Values for the Harmonic *N* of Data Points with Two Samples Using Two Different Formulas.

Size of Group 1 (*x*1)	Size of Group 2 (*x*2)	Formula 1 $\frac{N}{\frac{1}{x1}+\frac{1}{x2}+\cdots\frac{1}{xn}}$	Formula 2 $\frac{N\cdot N}{\frac{1}{x1}+\frac{1}{x2}+\cdots\frac{1}{xn}}$
15	15	15	30
500	500	500	1000
100	900	180	360

N is the number of groups in the comparison and *xi* is the size of each individual group.

Te Nijenhuis and van der Flier [34] used the formula $\frac{N\cdot N}{\frac{1}{x1}+\frac{1}{x2}+\cdots\frac{1}{xn}}$ where, again, *N* is the number of groups and x_i is the size of each individual group. For a datapoint with samples of 100 and 900, the value of the harmonic *N* then becomes 360, which is quite conservative, but not as strict as the value of only 180 for the first formula (see Table 2). For data points with samples of 15 and 15, the total sample size now becomes 30, and for a datapoint with samples of 500 and 500 the total sample size now becomes 1000 (see Table 2), which is in line with the reasoning in Hunter and Schmidt [39] mentioned above. We therefore continue to use this formula, which is based on sound reasoning, namely that data points consisting of samples with widely differing *N*s receive a substantially reduced weight in a meta-analysis, and that data points based on samples with highly comparable weights receive a weight based on the total number of research participants in these samples.

3. Results

The results of the studies on the vector correlation between *g* loadings and the score differences between adult Blacks and Whites (*d*) are shown in Table 3. The table shows data derived from 10 studies,

yielding 15 data correlations, with participants numbering a total of 251,085 Whites, 22,326 Blacks, and with a total harmonic N = 76,884. It also lists the reference for the study, the cognitive ability test used, the vector correlation between g loadings and d, the mean age, and the age range. The correlations are positive in sign, and the large majority of them are substantial in magnitude. Table 4 presents the results of the bare bones meta-analysis of the 15 data points. Table 4 shows the number of correlation coefficients (K), total sample size (N), the mean-weighted vector correlation (mean r), and the standard deviation of the vector correlation (SD_r). The last column presents the percentage of variance explained by sampling error ($\%VE$). The analysis of all 15 data points yields a mean vector correlation of 0.57, with 0.6% of the variance in the observed correlations explained by sampling error. This percentage is very low and suggests the presence of a strong moderator or several moderators.

Table 3. Studies of Correlations between g Loadings and Adult Black/White Differences.

Study	Test	r ($d \times g$)	$N_{subtests}$	N_{Whites}	N_{Blacks}	$N_{harmonic}$	Mean Age [1] (Range)
Carretta [41]	AFOQT	0.56	16	212,238	12,647	47,743	21 (18–27)
Murray [42]	WJ-I	0.35	6	3329	436	1542	(6–65)
	WJ-II	0.50	7	3573	807	2633	(6–65)
	WJ-III	0.72	7	2592	426	1463	(6–65)
Kaufman *et al.* [43]	WAIS-R	0.59	11	344	50	175	(16–19)
		0.67	11	440	50	180	(20–34)
		0.64	11	443	51	183	(35–54)
		0.48	11	437	41	150	(55–74)
Sternberg [44]	Various	0.46	11	348	47	83	(18–22) [1]
Department of Defense [45] [3]	ASVAB	0.30	10	5533	2298	6495	19.5 [2] (16–23)
Department of Labor [46] [3]	GATB Aptitudes	0.71	8	4001	2416	6025	40 (16–70)
Hennessy and Merrifield [47] [3]	CGP	0.66	10	1818	431	1394	18 (17–19)
National Longitudinal Study [3,4]	CGP, SAT, ACT	0.78	12	12,275	1938	6695	18 (16–23)
Nyborg and Jensen [38]	Various	0.81	16	3535	502	1758	19.9 (17–25)
Veroff *et al.* [48] [3]	Various	0.46	6	179	186	365	(18–49)

$N_{harmonic}$ is computed using the formula $\frac{4}{\frac{1}{n1}+\frac{1}{n2}}$ where $n1$ and $n2$ are the amount of participants in group $n1$ and $n2$, respectively. [1] Mean age not known for all groups; [2] Estimated; [3] These studies were taken from Jensen [37]; [4] Reference not given in Jensen [37]. AFOQT: Air Force Officer Qualifying Test; WJ-I/II/III: Woodcock-Johnson I/II/III; WAIS-R: Wechsler Adult Intelligence Scale—Revised; ASVAB: Armed Services Vocational Aptitude Battery; GATB: General Aptitude Test Battery; CGP: Comparative Guidance and Placement Program's test battery; SAT: Scholastic Aptitude Test; ACT: American College Testing.

Table 4. Exploratory Bare Bones Meta-analytical Results for Correlations between g Loadings and Adult Black/White Differences.

K	N_h	Mean r	SD_r	$\%VE$
15	76,884	0.57	0.12	0.6

Bare bones meta-analytical results: Score differences between adult Blacks and Whites, and g loadings. K = number of correlations; N = total sample size; mean r = mean-weighted vector correlation; SD_r = standard deviation of observed correlation; $\%VE$ = percentage of variance accounted for by sampling errors.

4. Discussion

We meta-analytically tested Spearman's hypothesis on Black and White adults. Spearman's hypothesis was already confirmed for children and adolescents in a large number of studies, and is

now confirmed comparing Black and White adults as well. The meta-analytical sample-size weighted correlation of *rho* = 0.57 we found for adults is highly similar to the mean correlation found in Jensen [1], who reported a correlation of $r = 0.59$. The differences between Black and White adults are very strongly in line with those previously found for children and adults.

We end on a cautionary note concerning conditions that are not fulfilled in our study, making our conclusions only conditionally valid. To the best of our knowledge, MGCFA has not been used for testing Spearman's hypothesis in the last decade, but how often a method has been used is not an evaluation criterion for truth finding. Measurement invariance is, strictly speaking, a necessary condition on a priori grounds. The fact that in the present study we did not have access to the original datasets means that we simply could not test for measurement invariance, so it is possible that some of the datasets when analyzed using MGCFA would have shown a lack of measurement invariance to a certain degree. Moreover, although some of the data points in our meta-analysis come from studies by Jensen, Jensen's use of congruence coefficients, for instance, does not prove measurement invariance. Indeed, in the other data points in our meta-analysis we do not carry out the statistical analyses suggested by Jensen [7], and this is an additional reason for a cautionary note.

Although it is good research practice to aim for the best method, we repeat that we employed a trade-off where we collected a substantial number of studies that could only be analyzed using non-optimal statistical techniques, but which allowed the use of the powerful technique of meta-analysis. For a well-argued trade-off leading to the inclusion of many studies of lesser methodological quality allowing a huge meta-analysis, we refer the interested reader to the meta-analysis on the effects of organizational development by Rodgers and Hunter [50].

Acknowledgments: We like to thank two anonymous reviewers and Paul de Boeck for their helpful and constructive feedback.

Author Contributions: Jan te Nijenhuis conceptualized and designed the study and drafted Introduction. Jan te Nijenhuis and Michael van den Hoek collected and analyzed the data, and drafted Method, Results, and Discussion.

Conflicts of Interest: The authors declare no conflict of interest.

References

References marked with an asterisk were included in the meta-analysis.

1. Jensen, A.R. *The g Factor: The Science of Mental Ability*; Praeger: Westport, CT, USA, 1998.
2. Jensen, A.R. *Bias in Mental Testing*; Free Press: New York, NY, USA, 1980.
3. Te Nijenhuis, J.; Al-Shahomee, A.A.; van den Hoek, M.; Allik, J.; Grigoriev, A.; Dragt, J. Spearman's hypothesis tested comparing Libyan secondary school children with various other groups of secondary school children on the items of the Standard Progressive Matrices. *Intelligence* **2015**, *50*, 118–124. [CrossRef]
4. Jensen, A.R.; Weng, L.J. What is a good *g*? *Intelligence* **1994**, *18*, 231–258. [CrossRef]
5. Gottfredson, L.S. Why *g* matters: The complexity of everyday life. *Intelligence* **1997**, *24*, 79–132. [CrossRef]
6. Spearman, C. *The Nature of "Intelligence" and the Principles of Cognition*; Macmillan: London, UK, 1923.
7. Jensen, A.R. Spearman's hypothesis tested with chronometric information-processing tasks. *Intelligence* **1993**, *17*, 47–77. [CrossRef]
8. Whetzel, D.L.; McDaniel, M.A.; Nguyen, N.T. Subgroup differences in situational judgment test performance. *Hum. Perform.* **2008**, *21*, 291–309. [CrossRef]
9. Goldstein, H.W.; Yusko, K.P.; Braverman, E.P.; Smith, D.B.; Chung, B. The role of cognitive ability in the subgroup differences and incremental validity of assessment center exercises. *Pers. Psychol.* **1998**, *51*, 357–374. [CrossRef]
10. Goldstein, H.W.; Yusko, K.P.; Nicolopoulos, V. Exploring Black-White subgroup differences of managerial competencies. *Pers. Psychol.* **2001**, *54*, 783–807. [CrossRef]
11. Rushton, J.P. Jensen effects and African/Coloured/Indian/White differences on Raven's Standard Progressive Matrices in South Africa. *Personal. Individ. Differ.* **2002**, *33*, 1279–1284. [CrossRef]

12. Rushton, J.P.; Čvorović, J.; Bons, T.A. General mental ability in South Asians: Data from three Roma (Gypsy) communities in Serbia. *Intelligence* **2007**, *35*, 1–12. [CrossRef]
13. Rushton, J.P.; Skuy, M. Performance in Raven's Matrices by African and White university students in South Africa. *Intelligence* **2000**, *28*, 251–265. [CrossRef]
14. Rushton, J.P.; Skuy, M.; Fridjhon, P. Jensen effects among African, Indian, and White engineering students in South Africa on Raven's Standard Progressive Matrices. *Intelligence* **2002**, *30*, 409–423.
15. Rushton, J.P.; Skuy, M.; Fridjhon, P. Performance on Raven's Advanced Progressive Matrices by African, East Indian, and White engineering students in South Africa. *Intelligence* **2003**, *31*, 123–137. [CrossRef]
16. Diaz, A.; Sellami, K.; Infanzón, E.; Lynn, R. A comparative study of general intelligence in Spanish and Moroccan samples. *Span. J. Psychol.* **2012**, *15*, 526–532. [CrossRef] [PubMed]
17. Te Nijenhuis, J.; Al-Shahomee, A.A.; van den Hoek, M.; Grigoriev, A.; Repko, J. Spearman's hypothesis tested comparing Libyan adults with various other groups of adults on the items of the Standard Progressive Matrices. *Intelligence* **2015**, *50*, 114–117. [CrossRef]
18. Te Nijenhuis, J.; Bakhiet, S.F.; van den Hoek, M.; Repko, J.; Allik, J.; Žebec, M.S.; Sukhanovskiy, V.; Abduljabbar, A.S. Spearman's hypothesis tested comparing Sudanese children and adolescents with various other groups of children and adolescents on the items of the Standard Progressive Matrices. *Intelligence* **2016**, *56*, 46–57. [CrossRef]
19. Te Nijenhuis, J.; Grigoriev, A.; van den Hoek, M. Spearman's hypothesis tested in Kazakhstan on items of the Standard Progressive Matrices Plus. *Personal. Individ. Differ.* **2016**, *92*, 191–193. [CrossRef]
20. Ganzach, Y. Another look at Spearman's hypothesis and relationship between Digit Span and General Mental Ability. *Learn. Individ. Differ.* **2016**, *45*, 128–132. [CrossRef]
21. Jensen, A.R.; Figueroa, R.A. Forward and backward digit span interaction with race and IQ: Predictions from Jensen's theory. *J. Educ. Psychol.* **1975**, *67*, 882–893. [CrossRef] [PubMed]
22. David, H. Response to: Another look at the Spearman's hypothesis and relationship between digit span and General Mental Ability. *Learn. Individ. Differ.* **2016**, *45*, 133–134. [CrossRef]
23. Ganzach, Y. On general mental ability, digit span and Spearman's hypothesis. *Learn. Individ. Differ.* **2016**, *45*, 135–136. [CrossRef]
24. Ashton, M.C.; Lee, K. Problems with the method of correlated vectors. *Intelligence* **2005**, *33*, 431–444. [CrossRef]
25. Dolan, C.V. Investigating Spearman's hypothesis by means of multi-group confirmatory factor analysis. *Multivar. Behav. Res.* **2000**, *35*, 21–50. [CrossRef] [PubMed]
26. Woodley, M.A.; te Nijenhuis, J.; Must, O.; Must, A. Controlling for increased guessing enhances the independence of the Flynn effect from *g*: The return of the Brand effect. *Intelligence* **2014**, *43*, 27–34. [CrossRef]
27. Schmidt, F.L.; Hunter, J.E. *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*, 3rd ed.; Sage: Beverly Hills, CA, USA, 2015.
28. Te Nijenhuis, J.; Jongeneel-Grimen, B.; Armstrong, E. Are adoption gains on the *g* factor? A meta-analyis. *Personal. Individ. Differ.* **2015**, *73*, 56–60. [CrossRef]
29. Te Nijenhuis, J.; Kura, K.; Hur, Y.M. The correlation between g loadings and heritability in Japan: A meta-analysis. *Intelligence* **2014**, *46*, 275–282. [CrossRef]
30. Woodley, M.A.; Fernandes, H.B.F.; te Nijenhuis, J. Differences in cognitive abilities among primates are concentrated on *G*: Phenotypic and phylogenetic comparisons with two meta-analytical databases. *Intelligence* **2014**, *46*, 311–322.
31. Te Nijenhuis, J.; Jongeneel-Grimen, B.; Kierkegaard, E.O.W. Are Headstart gains on the *g* factor? A meta-analysis. *Intelligence* **2014**, *46*, 209–215. [CrossRef]
32. Te Nijenhuis, J.; David, H.; Metzen, D.; Armstrong, E.L. Spearman's hypothesis tested on European Jews vs non-Jewish Whites and vs Oriental Jews: Two meta-analyses. *Intelligence* **2014**, *44*, 15–18. [CrossRef]
33. Flynn, J.R.; te Nijenhuis, J.; Metzen, D. The *g* beyond Spearman's *g*: Flynn's paradoxes resolved using four exploratory meta-analyses. *Intelligence* **2014**, *42*, 1–10. [CrossRef]
34. Te Nijenhuis, J.; van der Flier, H. Is the Flynn effect on *g*? A meta-analysis. *Intelligence* **2013**, *41*, 802–807. [CrossRef]
35. Van der Linden, D.; te Nijenhuis, J.; Bakker, A. The General Factor of Personality: A meta-analysis of Big Five intercorrelations and a criterion-related validity study. *J. Res. Personal.* **2010**, *44*, 315–327. [CrossRef]

36. Te Nijenhuis, J.; van Vianen, A.; van der Flier, H. Score gains on *g*-loaded tests: No *g*. *Intelligence* **2007**, *35*, 283–300. [CrossRef]
37. * Jensen, A.R. The nature of Black-White differences on various psychometric tests: Spearman's hypothesis. *Behav. Brain Sci.* **1985**, *8*, 193–263. [CrossRef]
38. * Nyborg, H.; Jensen, A.R. Black-white differences on various psychometric tests: Spearman's hypothesis tested on American armed services veterans. *Personal. Individ. Differ.* **2000**, *28*, 593–599. [CrossRef]
39. Hunter, J.E.; Schmidt, F.L. *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*, 2nd ed.; Sage: Thousand Oaks, CA, USA, 2004.
40. *Hunter & Schmidt Meta-Analysis Programs*; version 1.1; University of Iowa, Department of Management & Organization: Iowa City, IA, USA, 2004.
41. * Carretta, T.R. Group differences on US Air Force pilot selection tests. *Int. J. Sel. Assess.* **1997**, *5*, 115–127. [CrossRef]
42. * Murray, C. The magnitude and components of change in the Black-White IQ difference from 1920 to 1991: A birth cohort analysis of the Woodcock-Johnson standardizations. *Intelligence* **2007**, *35*, 305–318. [CrossRef]
43. * Kaufman, A.S.; McLean, J.E.; Reynolds, C.R. Sex, race, residence, region, and education differences on the 11 WAIS-R subtests. *J. Clin. Psychol.* **1988**, *44*, 231–247. [CrossRef]
44. * Sternberg, R.J. The Rainbow Project: Enhancing the SAT through assessment of analytical, practical, and creative skills. *Intelligence* **2006**, *34*, 321–350. [CrossRef]
45. * U.S. Department of Defense. *Profile of American Youth: 1980 Nationwide Administration of the Armed Services Vocational Aptitude Battery*; Office of the Assistant Secretary of Defense (Manpower, Reserve Affairs, and Logistics): Washington, DC, USA, 1982.
46. * U.S. Department of Labor, Manpower Administration. *Manual for the USES General Aptitude Test Battery*; U.S. Department of Labor, Manpower Administration: Washington, DC, USA, 1970.
47. * Hennessy, J.J.; Merrifield, P.R. A comparison of the factor structures of mental abilities in four ethnic groups. *J. Educ. Psychol.* **1976**, *68*, 754–759. [CrossRef]
48. * Veroff, J.; McClelland, L. *Measuring Intelligence and Achievement Motivation in Surveys. Final Report to U.S. Department of Health, Education and Welfare, Office of Economic Opportunity: Contract No. OEO-4180*; Survey Research Center, Institute for Social Research, University of Michigan: Ann Arbor, MI, USA, 1971.
49. Klockars, A.J.; Sax, G. *Multiple Comparisons*; Sage: Newbury Park, CL, USA, 1987.
50. Rodgers, R.; Hunter, J.E. The methodological war of the "Hardheads" *versus* the "softheads". *J. Appl. Behav. Sci.* **1996**, *32*, 189–208. [CrossRef]