psych

MDPI

Review

A Meta-Analysis of Spearman's Hypothesis Tested on Latin-American Hispanics, Including a New Way to Correct for Imperfectly Measuring the Construct of *g*

Jan te Nijenhuis *, Michael van den Hoek and Joep Dragt

Work and Organizational Psychology, University of Amsterdam, Nieuwe Achtergracht 129B, 1018 WS Amsterdam, The Netherlands; Michael.v.den.Hoek@gmail.com (M.v.d.H.); JoepDragt@hotmail.com (J.D.)

* Correspondence: JanteNijenhuis@planet.nl

Received: 9 March 2019; Accepted: 12 April 2019; Published: 18 April 2019

Abstract: Spearman's hypothesis states that the difference in intelligence between groups is a function of the *g* loadings of the subtests, where larger differences are found on tests with higher *g* loadings. This finding has consistently been supported on various groups. In this study we look at samples of Latin-American Hispanics in comparison to Whites. We carried out a meta-analysis based on 14 data points and a total of 16,813 Latin-American Hispanics, including a new way to correct for imperfectly measuring the construct of *g*. Spearman's hypothesis was strongly supported with a mean *r* of 0.63. After correction for various statistical artifacts this value became *rho* = 0.91. Therefore, we conclude that Spearman's hypothesis also holds true for White/Latin-American Hispanic differences.

Keywords: Spearman's hypothesis; Latin-American Hispanics; IQ; *g* loadings; group differences

1. Introduction

The US Census Bureau states that Hispanics are those people who classify themselves as "Mexican", "Puerto Rican", or "Cuban", as well as those whose origins are from Spain, the Spanish-speaking countries of Central or South America, or the Dominican Republic. Origin can be seen as the country of birth, lineage, nationality group, or heritage of the person or the person's parents or ancestors before their arrival in the United States. The U.S. Census Bureau also states that people who identify their origin as Hispanic may be of any race [1]. The Census Bureau does not classify persons of Portuguese or Brazilian descent as Hispanic [2]. The term Hispanic is linked to the name of the former Roman province of Hispania (from which we derive the modern name of Spain); present-day Portugal belonged to the old Roman province of Lusitania, not Hispania. Hispanics are a relatively large group in the United States and account for approximately 15.1% of the US population [3].

The term Hispanic is used to denote an ethnicity rather than a race, so someone of Hispanic descent could racially be White, Black, Amerindian, or Asian, and therefore there are large differences between the various groups of Hispanics. It is a well-known fact that the gene pool of Mexico is approximately 80% Amerindian. This figure is even higher in southern Mexico and in most of the Central-American countries, from which most Hispanic-Americans come today. These Hispanics come heavily from the lower classes—raising the issue of selective migration that probably makes them unrepresentative of their home group. For example, given the class structure of these countries, lower-class migrants will be more Indian than their countries as a whole. It has been estimated that the Mexican-American Hispanic gene pool is 90% Amerindian. Puerto-Rican Hispanics, by contrast, are often of African descent and few if any are of Amerindian decent. These Hispanics also come mostly from the lower classes—again raising the issue of selective migration and unrepresentative samples.

In contrast, Cuban-American Hispanics are essentially entirely of European decent, and they come mostly from the middle and upper-middle classes of Cuba, having been forced out by Castro

or having fled the communist government. A clear exception was formed by a small percentage of Cuban criminals released from prison among the mass emigration of Cubans during the so-called 'Mariel boatlift' in 1980. Again, this indicates the question of selective migration, with (in this case) Cuban-American Hispanics being educationally and intellectually superior on average to the population of Cubans as a whole, to other Hispanics in the U.S., and possibly even to the White U.S. population.

In the present paper we excluded Spanish nationals as our focus is on groups that generally have a substantially lower mean IQ than Anglo-Americans, so the generally lower IQ needs explaining, or they may have a mean IQ that is quite similar to or possibly even higher than that of Anglo-Americans, but a different cultural background; the Spanish mean IQ is very similar to the U.S. Anglo-American mean IQ [4]. From a practical point of view, to the best of our knowledge, there are simply no studies comparing the scores of Spanish nationals and Anglo-Americans. We feel that excluding Spanish nationals leads to a homogeneous study group in our meta-analysis, in the sense that all these populations are expected to have lower mean IQ scores than Anglo-Americans. So, our focus is on Latin-American Hispanics.

Latin-American Hispanics often do not perform well in educational settings. They have a lower level of educational attainment than Whites and they show a much higher drop-out rate at school than Whites; this effect is even stronger for Latin-American Hispanics born outside of the US [5]. Furthermore, they are less likely to obtain a bachelor's or a higher-level degree than Whites, with only about 13% of Latin-American Hispanics obtaining such a degree, compared to 31% of Whites [6]. This discrepancy is sometimes blamed on Latin-American Hispanics having relatively poor English-language skills: they often speak Spanish at home and some never learned English at all. Indeed, research shows that in 2012 13.1% of the Hispanic population ages 5–17 did not speak English very well (11.5% for U.S.-born; 29.8% for foreign-born) and ~39% of the Hispanic population 18 and up does not speak English very well (10.6% for U.S.-born; 68.2% for foreign-born) [7].

Another possible cause of the poor educational performance is the below-average intelligence of Latin-American Hispanics. Lynn [8] used the Differential Ability Test and showed an IQ gap of about nine points between White and Hispanic children. A meta-analysis by Roth, Bevier, Bobko, Switzer, and Tyler [9] found that Hispanic adults perform more poorly on tests of intelligence in the workplace when compared to White adults.

Spearman's hypothesis [10] states that group differences on standardized intelligence test batteries between groups of Blacks and Whites can be easily explained: there are larger differences on the more cognitively complex subtests and smaller differences on the less cognitively complex subtests. Jensen devised the method of correlated vectors as the test of Spearman's hypothesis: the *g* loadings of all the subtests are computed to construct the vector of *g* loadings and the standardized group differences on the same subtests are computed to construct the *d* vector, and then these two are correlated. Within the U.S., there is a lot of empirical support for Spearman's hypothesis comparing Whites and Blacks [11], Amerindians [12], and Jews [13]. It should be noted that European Jews have higher mean scores than both non-Jewish Whites and Oriental Jews. Following Jensen's logic of subtracting the IQ scores of the lower-scoring group from the IQ scores of the higher-scoring group, the scores of the non-Jewish Whites were subtracted from the scores of the European Jews, and the scores of Oriental Jews were subtracted from the score of the European Jews. However, so far Spearman's hypothesis has not been extensively tested in comparisons of Latin-American Hispanics and Whites. Support for Spearman's hypothesis tested on Latin-American Hispanics has been found by Hartmann, Kruuse, and Nyborg [14]. Whites scored ~0.8 standard deviations higher than Latin-American Hispanics in two different samples, with the first sample based on data from the Centre for Disease Control and the second sample based on data from the National Longitudinal Survey of Youth (NLSY79) in the U.S. In both samples they found that cognitive complexity of subtests explained the magnitude of group differences on the same subtests. Kane [15] tested Spearman's hypothesis on Latin-American Hispanics using the Universal Nonverbal Intelligence Test (UNIT), and Dalliard [16] tested Spearman's hypothesis using the Differential Ability

Scales-II (DAS-II); in both instances there was support. Ganzach (2016ab) [17,18] takes us back to Jensen's (Jensen & Figueroa, 1975) [19] work on Black/White differences before he started testing Spearman's hypothesis. In this early work, Jensen showed these differences were much smaller on the Wechsler subtest Digit Span Forward than on the Wechsler subtest Digit Span Backward, whereas the former subtest had higher complexity than the latter subtest. Ganzach (2016a) [17] used a nationally representative, large dataset to empirically check whether Jensen's findings on Black/White differences could be generalized to Hispanic/White differences; he found a reverse pattern for Hispanics: larger differences on Digit Span Forward and smaller differences on Digit Span Backward. Spearman's hypothesis cannot be tested on just two subtests, but the pattern in the Hispanic/White differences is opposite to that predicted by Spearman's hypothesis. This leads to the question how strongly Spearman's hypothesis will be confirmed in a meta-analysis. So, more research is clearly needed.

We chose to analyze our data with the method of correlated vectors (MCV). The MCV has been criticized by various researchers [20–22] and is considered controversial by some. We refer the interested reader to recent, detailed discussions on this topic (Woodley, te Nijenhuis, Must, & Must, 2014 [23]; te Nijenhuis & van den Hoek, 2016; [24] te Nijenhuis, Choi, van den Hoek, Valueva, & Lee, in press [25]). As an aside, we notice that the present study is on Spearman's hypothesis tested on subtests of an IQ battery, but that there is a recent discussion on the test of Spearman's hypothesis at the item level [Wicherts, 2018 [26]; te Nijenhuis & van den Hoek, 2018 [27]]; however, this discussion has little relevance for the present paper [25].

The goal of the current study is to test and extend Spearman's hypothesis using meta-analysis. This is especially important since Latin-American Hispanics are now the largest minority group in the US, and will likely become a more substantial part of the workforce. We expect to find similar results as found in the study by Hartmann et al., namely that the differences between Latin-American Hispanics and Whites on subtests will correlate with the *g* loading of the subtests.

2. Methods

2.1. Meta-Analysis

The purpose of this study is to determine whether the correlation between the *g* loadings and difference scores on IQ subtest between Whites and Latin-American Hispanics is strong and positive in sign. We carried out a meta-analysis where we tested Spearman's hypothesis in comparisons of Whites and Latin-American Hispanics in the U.S. We did not use Jensen's [28] procedure for testing Spearman's hypothesis, but a more simplified procedure [10], which does not include testing for measurement invariance. We carried out a full-fledged psychometric meta-analysis where we corrected for various statistical artifacts [29] using the software package developed by Schmidt and Le [30].

2.2. Rules for Inclusion

For studies to be included in a meta-analysis on Latin-American Hispanics three criteria had to be met. First, the IQ test had to be well-validated. Second, in order to obtain a reliable estimate of the correlation between each of the variables and *g* loadings, the cognitive batteries had to be based on a minimum of five subtests. Jensen [10,31] used six subtests as a minimum, but in recent meta-analyses using the method of correlated vectors [12,13,32–35] it has been empirically shown that if the underlying relationship is strongly positive or strongly negative a test battery with five subtests in the majority of cases still strongly bring out the theoretically expected correlation. However, in many cases using four subtests leads to unstable outcomes, although for exploratory purposes four subtests could still be used. In contrast, when the theoretically expected effect is not strongly positive or negative at least seven subtests are necessary for a reliable measurement [36,37]. Third, the samples had to be Hispanic, but we excluded Spanish nationals as our focus is on groups that generally have a lower mean IQ than Anglo-Americans, so the generally lower IQ needs explaining, or they may have a mean IQ that is quite similar to or possibly even higher than that of Anglo-Americans, but a different

cultural background; the Spanish mean IQ is very similar to the U.S. Anglo-American mean IQ [4]. From a practical point of view, to the best of our knowledge there are simply no studies comparing the scores of Spanish nationals and Anglo-Americans. We feel that excluding Spanish nationals leads to a homogeneous study group in our meta-analysis, in the sense that all these populations are expected to have lower IQ scores than Anglo-Americans. A sample by Sternberg et al. [38] was described as including Latinos, so it might include Portuguese and Brazilians and we therefore did not include it in the meta-analysis.

2.3. Searching and Screening Studies

Digital searches were carried out in online databases: Google Scholar, ProQuest, PsycINFO, and CataloguePlus (Primo). Terms used in the digital search were 'Hispanic', 'Latino', 'Chicano', 'Latin American', 'Mexico', 'Puerto Rico', and 'Cuba', as well as the names of all Central and South American countries. These terms were combined with the terms 'intelligence', 'IQ', 'mental ability', 'mental capacity', 'cognitive ability', 'aptitude', 'competence', 'differences', 'WISC', 'WAIS', 'WPPSI', 'K-ABC', and 'Woodcock Johnson'. We also used Spanish keywords to obtain more relevant data for our analysis: 'inteligencia', 'intelecto' 'capacidad cognitiva', and 'capacidad mental'. Further searches were carried out by following up on the references given in the studies obtained through digital means, as well as references of the studies of Hispanic intelligence reported by Lynn [39] (p. 108). These searches resulted in a total of 13 studies that were usable in our analysis, with a total of 14 data points.

2.4. g Loadings

To test Spearman's hypothesis, Pearson correlations between difference scores of the subtests of the IQ battery and *g* loadings were computed. In general, *g* loadings were computed by conducting a principal axis factor analysis on the correlation matrix of a test battery's subtest scores. The subtest's loadings on the first unrotated factor indicates the subtest's loading on *g*. *g* loadings were always matched to the age range of the groups involved in the comparison as close as possible. If the age range of the comparison groups comprised more than one age group of the IQ battery, we computed weighted-average *g* loadings of all age groups of the IQ battery that fall within the age range of the comparison groups. Whenever possible we used the *g* loadings from the largest available group for our analyses, such as standardization samples.

2.5. Calculating Glass' d

Since the White groups scored higher on most tests, the Glass' *d*s [40,41] were calculated by taking the scores of the Latin-American Hispanic group and subtracting them from the scores of the largest available White comparison group (mostly standardization samples when available). Referral is the act of officially sending someone to a person or authority that is authorized or better qualified to deal with them and referral groups were compared to other referral groups for a fair comparison. Then the results were divided by the standardization sample *SD* or the *SD* of the largest White sample available. Please note that our effect size is Glass' *d* because we did not use the pooled *SD*, but the *SD* of a White sample.

For the statistical formula for sampling error to apply, it could be argued that the denominator in calculating the Glass' *d* statistic should be the *N*-weighted average *SD* across the two groups, because if only one *SD* is used, the formula for sampling error variance is not known. However, in this case this would not make a lot of difference, since the White group is always much larger and so would strongly dominate the weighted average.

2.6. Studies Supplying Multiple Effect Sizes

Schmidt and Hunter [29] (p. 449) state that many researchers routinely compute effect sizes separately by sex and race, even when there is usually no good reason to expect that they will act as a moderator. Especially when the total sample is not large, the analysis of subgroups exacts a price

by increasing sampling error. Schmidt and Hunter [29] (p. 451) advised that the major cumulative analysis should be carried out with total-group effect sizes when the demographic variable has been shown to have little or no moderating effect when the total group effect size cannot be computed an average effect size can also be used [29] (pp. 449–451).

The study by Hartmann et al. [14] used samples from two different databases, so it supplies two data points to the meta-analysis. The study by Kaufman, McLean, and Kaufman [42]) reports data on two different age groups. Age does not appear to act as a clear and strong moderator of Spearman's hypothesis tested on IQ batteries: te Nijenhuis and van den Hoek [35] presented a large meta-analysis of Black adults and showed that the outcomes are strongly in line with those found for Black children [10]. Flemmer and Roid [43] present data split up by three education levels, but education level has not been suggested as a clear and strong moderator in the literature on Spearman's hypothesis. As there is no empirical proof that moderators play a substantial role, we therefore combined the samples, as suggested by Schmidt and Hunter [29] (ch. 10, pp. 449–451).

2.7. Correcting for Sampling Error, Reliability of the g Vector, and Reliability of the Glass' d Vector

In many cases sampling error explains the majority of the variation between studies, so the first step in a psychometric meta-analysis is to correct the collection of effect sizes for differences in sample size between the studies. We use the number of tests in the IQ battery as the sample size. For example, take a study that yields means and *SD*s for 12 subtests of an IQ battery on two groups and also obtains data that allow the computation of *g* loadings, then this number of twelve subtests is the number of data points. Each of these 12 data points includes one value of a *g* loading and one value of the *d* statistic. The correlation between these two vectors or variables is based on $N = 12$.

It is a highly frequent finding in psychometric meta-analysis that, when powerful moderators are absent, sampling error explains at least half the variance between the data points. When using number of subtests as an indicator of sampling error, meta-analyses based on the method of correlated vectors yields this classic outcome. When using sample size as the indicator of sampling error, as was done in the earlier meta-analyses where the Method of Correlated Vectors-Psychometric Meta-Analytical Hybrid Model was used, just a very small percentage of variance was explained, even with a quite substantial number of data points in the study. Moreover, it is important in psychometric meta-analysis (Schmidt & Hunter, 2015) [29] that all the corrections for statistical artifacts are statistically independent. When using sample size as an indicator of sampling error, there is a correlation with the correction for the reliability of the second vector, but when using the number of subtests as an indicator for sampling error there is statistical independence. Please note that a somewhat comparable principle can be found in Jensen's classic studies of Spearman's hypothesis: the outcome of the test of significance of the correlation between vectors is based on the number of subtests in the IQ battery, and not on the sample size.

Also note that the number of participants in the study is not forgotten in the Method of Correlated Vectors-Psychometric Meta-Analytical Hybrid Model: the correction for the reliability of the second vector is based upon the size of the sample. So, sample size plays a role in the size of the meta-analytical correlation and in the computation of the amount of variance between the data explained by statistical artifacts.

The meta-analytical results in this paper would be necessarily biased were we to carry out a bare bones meta-analysis. A bare bones meta-analysis is one in which there is no correction for the downward biasing effect of measurement error in either of the measures being correlated or related. The measures being correlated in the present study are the *g* vector and the Glass' *d* vector. It is not possible to estimate *g* loadings in a *g* vector without error; the estimates contain measurement error. The amount of measurement error is revealed, amongst others, in the reliability coefficient of the *g* vector. As illustrated in Jensen [10] (ch. 10), this reliability is estimated by the correlation between *g* vectors estimated on two different groups of similar size and of similar background taking the same test battery. If this reliability is known, then it can be used to correct the correlations in the meta-analysis

for the downward bias created by the measurement error. Estimates of this reliability are presented in Jensen [10], in te Nijenhuis, van Vianen, and van der Flier [32], and in te Nijenhuis and van der Flier [36]. The same consideration applies to the between-group differences on the tests in the form of Glass' *d* values. Again, the smaller the sample sizes in two groups, the more measurement error there will be in these Glass' *d* values, and vice versa.

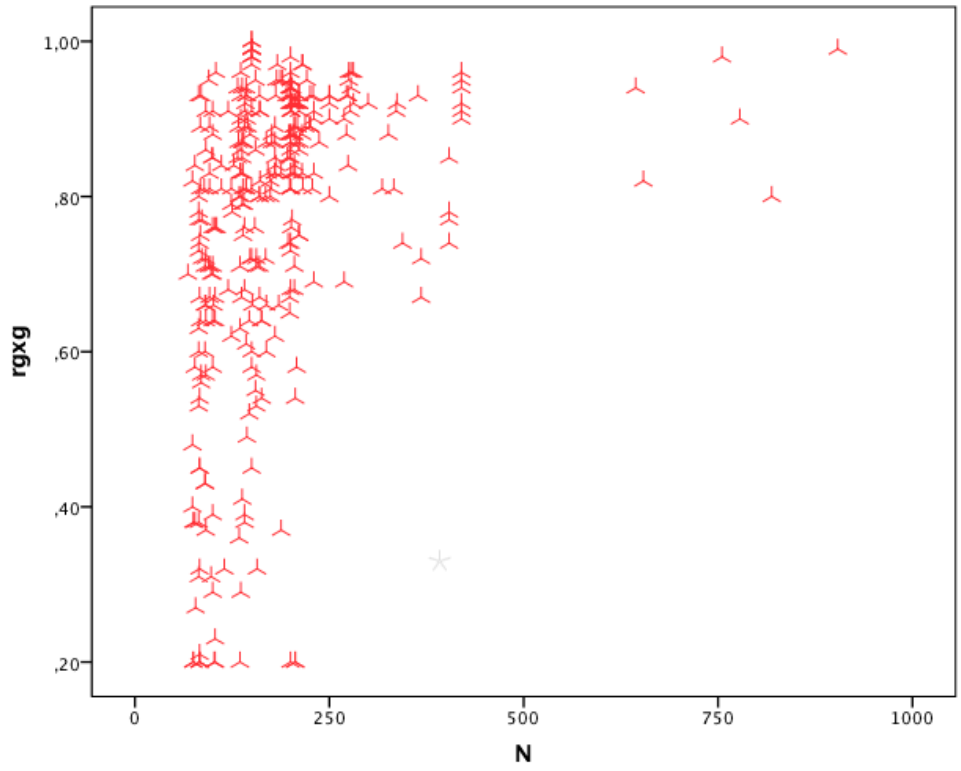
To correct for the reliability of the vector of *g* loadings we made use of the data reported in te Nijenhuis, Jongeneel-Grimen, and Armstrong [34], who constructed a distribution of reliabilities (see Figure 1). Several samples were compared that differed little on background variables. For the comparisons using children we chose samples that were highly comparable with regard to age. Samples of children in the age of 3 to 5 years were compared against other samples of children who did not differ more than 0.5 year of age. Samples of children between the ages of 6 and 17 years were compared against other samples of children who did not differ more than 1.5 year of age. For the comparisons of adults, samples of individuals between the ages of 18 and 95 years were compared.

Correlation matrices were collected from test manuals, books, articles, and technical reports. The large majority came from North America, with a large number of European countries, and also a substantial number from Korea, China, Hong Kong, and Australia. This resulted in ~700 data points, which led to 385 comparisons of *g* loadings of comparable groups which provided an indication of the reliability for that group [34].

If the items of a scale are measuring one construct, then the reliability estimates should be positive. But, if the scale is badly constructed, one could have a negative coefficient alpha, as the value is simply the average of all possible split-half correlations. In the case of intelligence batteries, there is more than a century of experience of constructing high-quality instruments, and all the subtests correlate highly with one another, so you would not expect negative reliabilities. In line with this, te Nijenhuis, Jongeneel-Grimen, and Armstrong [34] decided to smoothen their distributions by leaving out negative data points. In the present study we decided to use an additional technique to smoothen distributions namely leaving out extreme outliers, so data points with a reliability that differs very strongly from the general outcome for a specific sample size. Usually these extreme outliers indicate very low reliabilities, which would lead to overcorrection [29], and this is highly undesirable. Excluding data points indicating low reliabilities leads to less strong support of Spearman's hypothesis after corrections for statistical artifacts, so, the outcomes now become more conservative. Obviously, we were reluctant to remove data points.

It is generally difficult to describe a procedure for this process of outlier removal without arguably putting too much emphasis on statistics—we relied strongly on visual inspection and then only removed a possible outlier when it was a large number of *SD*s away from the other data points. We began with the visual inspection of the scatter plot of reliabilities against *N*s looking for extreme outliers. When we found them, we looked whether there was a cluster of data points to which they belonged. We then computed the standard deviation of the data points in the cluster, obviously without including the value of the reliability of the outliers. We note that the *SD* is computed on the values of the reliabilities. The next step was to compute the distance expressed in the *SD* of the reliabilities of the cluster between the mean value of the cluster and the outlier. When the distance was more than 10 *SD*s from the mean we regarded this as an extreme outlier and deleted it from the distribution. This led to five of the 385 data points being removed, which yields a very small percentage.

A scatter plot of reliabilities against *N*s showed that the larger *N* becomes, the higher the value of the reliability coefficients, with an asymptotic function between $r\ (g \times g)$ and *N*. However, because the extreme range on the x-axis resulted in a picture that is not informative, the regression line for $r\ (g \times g)$ and *N* is not reported. For the same reason we divided Figure 1 into three parts, each showing the scatter plot of reliability of the vector of *g* loadings and sample size for a specific range of *N*.



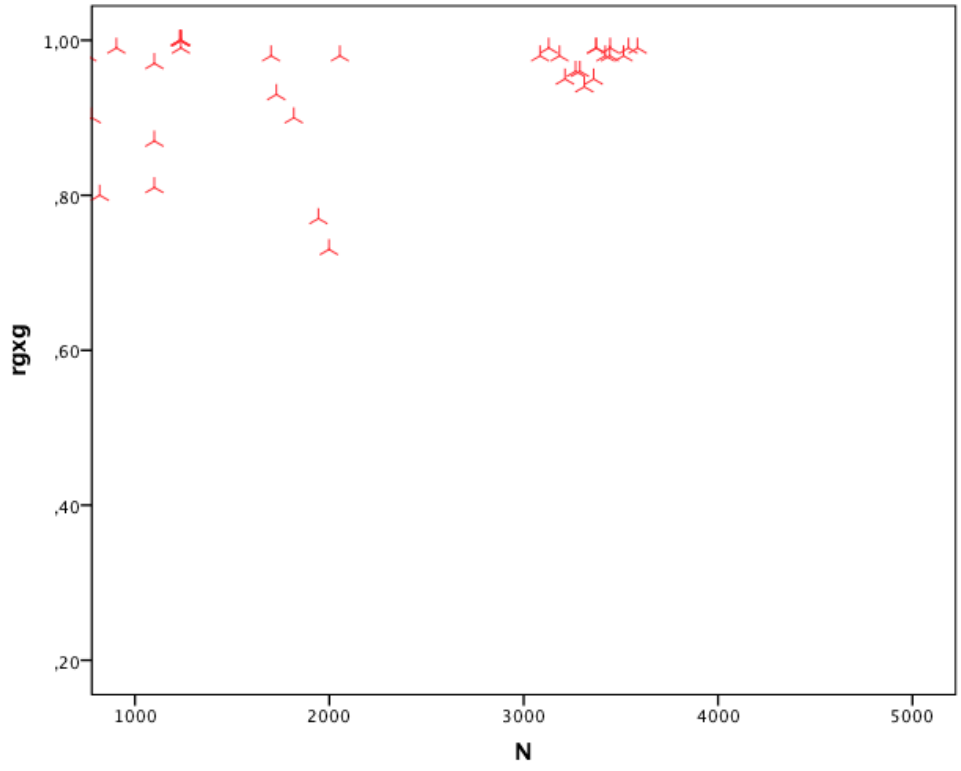


Figure 1. *Cont.*

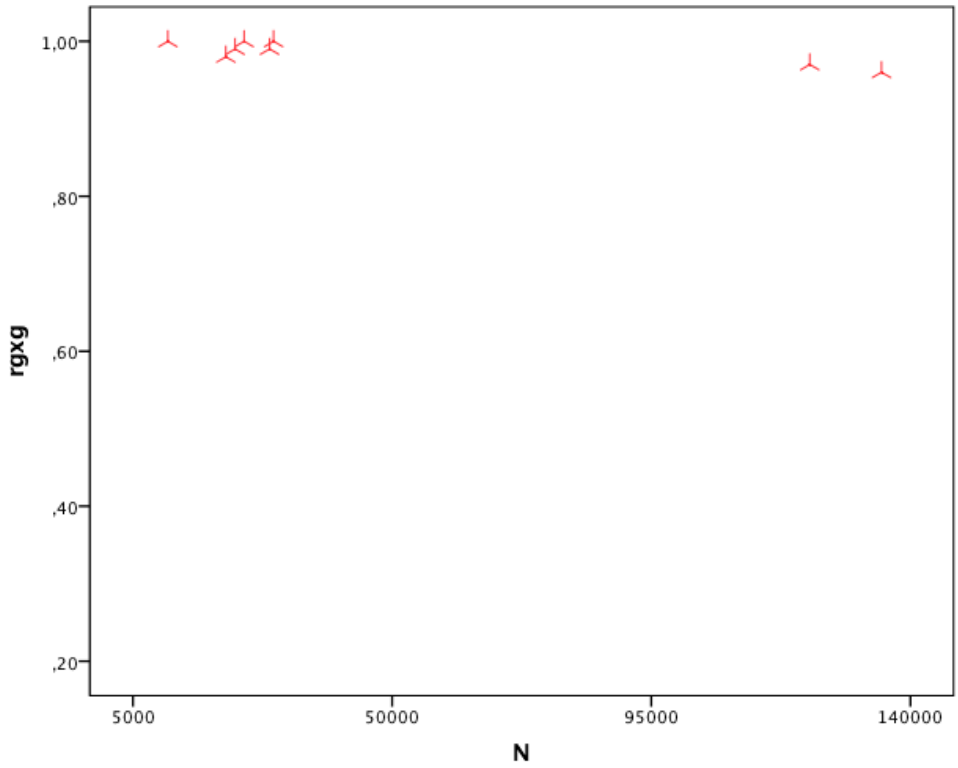


Figure 1. Three scatter plots of the reliability of the vector of *g* loadings and sample size each for different range of N.

The values of *r* (*g* × Glass' *d*) for Spearman's hypothesis tested on intelligence test batteries are attenuated by the reliability of the Glass' *d* vector for a given battery. When two samples have a comparable *N*, the average correlation between vectors is an estimate of the reliability of each vector. The reliability of the Glass' *d* vector was estimated using the present datasets, comparing samples that took the same test, and that differed little on background variables. For the comparisons using children, we chose samples that were highly comparable with regard to age, and for the comparisons of adults, we chose samples that were roughly comparable with regard to age. We only used comparisons that were comparable in size, background, and had similar populations (referral vs. nonreferral; see definition in Section 2.5). This yielded five possible comparisons. One comparison yielded a negative correlation, and therefore we did not include this comparison. Removing the negative correlation, we were left with four estimates of the reliability of the *d* vector. These reliabilities were plotted in Figure 2 and curves were fitted to the data points. As expected, the reliability of the Glass' *d* vector increases with sample size and a logistic curve was the best fit for the asymptotic function of *r* (Glass' *d* × Glass' *d*). Reliabilities for comparisons outside of the sample size range were estimated using the regression-line formula.

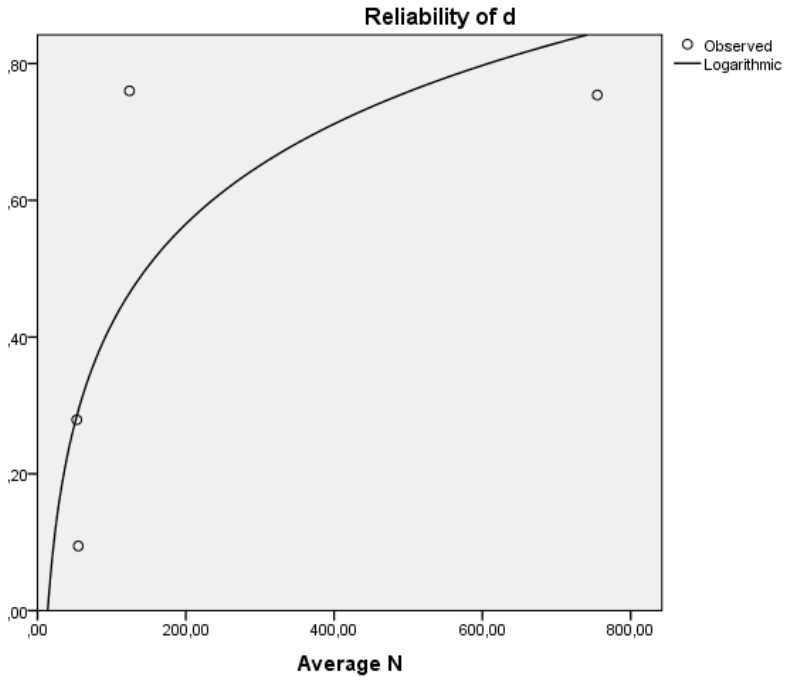


Figure 2. Reliability of r ($d \times d$) as a Function of Sample Size.

2.8. Correction for Restriction of Range in g Loadings

Jensen [10] (ch. 10) wrote that the values of r ($g \times$ Glass' d) are attenuated by the restriction of range of g loadings in many of the standard test batteries. The most highly g-loaded batteries tend to have the smallest range of variation in the subtests' g loadings. Jensen [10] (pp. 381–382) showed that restriction in the magnitude of g loadings strongly attenuates the correlation between g loadings and standardized group differences. Hunter and Schmidt [44] stated that the solution to variation in range is to define a reference population and express all correlations in terms of it. The Hunter and Schmidt meta-analytical program computes what the correlation in a given population would be if the standard deviation were the same as in the reference population. The standard deviations can be compared by dividing the standard deviation of the study population by the standard deviation of the reference group, that is, $u = SD_{\text{study}}/SD_{\text{ref}}$. To give an example, suppose the observed correlation is $r = 0.50$, the $SD_{\text{study}} = 0.150$, and the $SD_{\text{ref}} = 0.192$, then the value of u is 0.150/0.192 = 0.78, which results in a correction factor of 1.28, which yields a corrected correlation with a value of *rho* = 0.64.

Jensen [10] (p. 382) can be read as suggesting that the unrestricted value of *SD* can both be theoretically and empirically derived; he then goes on to empirically derive this value. Te Nijenhuis and coauthors [34] also chose to work with empirically-derived values of the unrestricted *SD*. As references they used tests that are broadly regarded as exemplary for the measurement of intelligence, namely the various versions of the Wechsler tests for children and adults. The average standard deviation of g loadings of the various versions of the Wechsler from datasets from countries all over the world was 0.132 for children and 0.107 for adults. They used these values as their reference in the studies with children and adults, respectively. In so doing, the *SD* of g loadings of all test batteries was compared to the average *SD* in g loadings in the Wechsler tests for children and adults, respectively.

In the present paper, however, we chose to work with theoretically-derived values of the unrestricted *SD*. Frank Schmidt (personal communication with Jan te Nijenhuis, 2010) suggested to use the maximal range of $g = 0$ to $g = 1$ for every meta-analytical test of Spearman's hypothesis, so not only for IQ batteries, but also for elementary cognitive tasks, safety suitability tests, school achievement tests, Situational Judgment Tests, and Assessment Center exercises. Flynn [45] (p. 36) also chose the theoretically-derived value of the unrestricted *SD*. The unrestricted population is defined as having g

loadings ranging from 0.001 to 0.999 with *SD* is 0.167. Flynn does not report details, but appears to use the values of $z = -3.00$ to $z = +3.00$ of a normal distribution equaling 6 *SD*s, thereby covering 99.9% of the distribution. In the present paper, we use the value of the unrestricted *SD* suggested by Flynn.

The Hunter and Schmidt meta-analytical program computes only the aforementioned four corrections for statistical artifacts. We will refer to the observed correlation corrected for sampling error, unreliability of the vector of *g* loadings, unreliability of the Glass' *d* vector, and range restriction as *rho*-4.

2.9. Correction for Deviation from Perfect Construct Validity

Jensen [10] (1ch. 10) writes that the more tests in a test battery and the higher their *g* loadings, the higher the *g* saturation of the composite score is. The Wechsler batteries have a large number of subtests with quite high *g* loadings, yielding a highly *g*-saturated composite score. Jensen [10] (pp. 90–91) states that the *g* score of the Wechsler batteries correlates more than 0.95 with the tests' IQ score. However, shorter batteries with a substantial number of tests with lower *g* loadings will lead to a composite with somewhat lower *g* saturation. Jensen [10] (ch. 10) states that the average *g* loading of an IQ score as measured by various standard IQ tests lies in the +0.80s. When this value is taken as an indication of the degree to which an IQ score is a reflection of "true" *g*, it can be estimated that a test's *g* score correlates ~0.85 with "true" *g*. As *g* loadings represent the correlations of tests with the *g* score, it is most likely that most empirical *g* loadings will underestimate "true" *g* loadings; therefore, empirical *g* loadings correlate about 0.85 with "true" *g* loadings. As the Schmidt and Le [30] computer program only includes corrections for the first four artifacts, the correction for deviation from perfect construct validity has to be carried out on the values of *r* (*g* × Glass' *d*) after correction for the first four artifacts. To limit the risk of overcorrection, in previous meta-analyses using the method of correlated vectors a conservative choice of the value of 0.90 for the correction was made [32,36]. The observed correlation after corrections for sampling error, unreliability, range restriction, and imperfect construct validity was referred to as rho-5, as it was corrected for five statistical artifacts.

Another way to estimate the distribution of values necessary for the fifth correction is based on using samples that took a large number of cognitive tests. Wechsler test data were analyzed using a formula given by Jensen [10] (pp. 103–104) to compute the *g*-loadedness of a sum score:

$$\{1 + \{\sum [r^2_{sg}/(1 - r^2_{sg})]^{-1}\}\}^{-0.5} \quad (1)$$

where r^2_{sg} = each subtest's squared *g* loading. The formula implies that longer test batteries in general are more *g*-loaded than shorter test batteries, with *g*-loadedness being an asymptotic function of the number of subtests. Using this formula resulted in a *g* loading of 0.92–0.95 for the various Wechsler Full Scale scores based on 10–12 subtests. It may be that having about fifteen subtests from one or more test batteries gives one a total score with perfect *g*-loadedness. The next step is to argue that when using datasets with many cognitive tests the larger the collection of subtests becomes, the more the resulting *g* score approaches Jensen's concept of "true" *g*. The final step is to compute a *g* score based on, for instance, six subtests from a large collection of cognitive tests and to correlate this *g* score with a *g* score based on, say, 25 subtests yielding an estimate of the correlation of the sum score based on six subtests with "true" *g*. Various combinations of six subtests from a larger collection are possible and their correlations with *g* based on a large number of subtests yield an estimate of the distribution of the value necessary for the correction for imperfectly measuring the construct of *g* when using a battery consisting of six cognitive tests.

We wanted to know how strongly the correlation between "true" *g* and the *g* from an artificial test battery is a function of the number of subtests. Therefore, we used a dataset with a large number of cognitive tests to create a distribution of the values of the correlation of a test battery with the construct "true" *g*. It was expected that the measurement of "true" *g* by a test battery was an asymptotic function of the number of subtests. Several analyses were carried out on the dataset.

2.9.1. Research Participants

Bleichrodt, Resing, Drenth, and Zaal [46] carried out a study using two IQ tests, namely, both the Dutch WISC-R and the RAKIT. The RAKIT was given to a nationally representative sample of 1415 Dutch children of seven age groups (4–11 years of age). The Dutch WISC-R was also taken by a subsample of 469 children aged 6–9 from 60 primary schools from the RAKIT sample, with two weeks in between the taking of tests. In 29 schools the WISC-R was taken first and then the RAKIT. In 31 schools the RAKIT was taken first, then the WISC-R.

2.9.2. Psychometric Variables

Each test battery consists of 12 subtests that are highly diverse in the types of abilities, information content, and cognitive skills they call for. The subtests of the RAKIT [46]:

1. Closure; the child is given very incomplete pictures and has to figure out the complete picture. According to Carroll's [47] taxonomy, this subtest is a measure of Closure Speed at stratum I, which makes this subtest a measure of Broad Visual Perception at stratum II.
2. Exclusion; out of four abstract figures the child has to select the one that is different from the other three. The child has to detect the necessary rule to solve the task. This subtest measures Induction at stratum I, which makes it a measure of Fluid Intelligence at stratum II.
3. Memory Span; the child has to memorize figures put on cards and the sequence in which they are presented. After five seconds the card is turned and the child has to reproduce the figures in the right sequence using blocks on which the figures are printed. The subtest contains a series with concrete figures and a series with abstract figures. Both series measure (Visual) Memory Span at stratum I. Both series fall under General Memory and Learning at stratum II.
4. Verbal Meaning; words are presented to the child in an auditory fashion and from four figures the child has to choose the one which resembles the word it has just heard. This subtest measures Lexical Knowledge at stratum I and is a measure of Crystallized Intelligence at stratum II.
5. Mazes; the child has to go through a maze with a stick as fast as they can. Because of the speed factor this subtest is a measure of Spatial Scanning at stratum I, which falls under Broad Visual Perception at stratum II.
6. Analogies; the child has to complete verbal analogies that are stated as follows: A:B is like C: . . . (there are four options to choose from). The constructors of this subtest tried to avoid measuring Lexical Knowledge, by including only those words that are highly frequently used in ordinary life. All words in the analogy items are accompanied by illustrations, so as to reduce the verbal aspect of the task to a minimum. This subtest is a measure of Induction at stratum I which makes it a measure of Fluid Intelligence at stratum II.
7. Quantity; in this multiple-choice test the child has to make comparisons between pictures, differing in volume, length, weight, and surface. This subtest is a measure of Quantitative Reasoning at stratum I, which measures Fluid Intelligence at stratum II.
8. Disks; the child has to use pins to put disks with two, three, or four holes on a board as fast as possible until three layers of disks are on the board. This subtest is a measure of Spatial Relations at stratum I, which measures Broad Visual Perception at stratum II.
9. Learning Names; the child has to memorize the names of different butterflies and cats using pictures presented on cardboard. This subtest measures Associative Memory at stratum I, which makes it a measure of General Memory and Learning at stratum II.
10. Hidden Figures; the child has to discover which of six figures is hidden in a complex drawing. This subtest is a measure of Flexibility of Closure at stratum I, which makes it a measure of Broad Visual Perception at stratum II.
11. Idea Production; the child has to name as many words, objects, or situations as possible that can be associated with a broad category within a certain time span, for example: "What can you eat?"

This subtest is a measure of Ideational Fluency at stratum I, which is a measure of Broad Retrieval Ability at stratum II.

12. Storytelling; the child has to tell as much as possible about a picture on a board and what could happen to the persons or objects in the picture. The total score is composed of both quantitative measures (number of words, number of relations, did or did not the child tell a plot, etc.) and qualitative measures (did the child grasp the central meaning of the story). This subtest consists of different elements and measures at stratum I: Naming Facility and Ideational Fluency, Sequential Reasoning, and to some extent Communication Ability. These stratum-I abilities are measures of Broad Retrieval Ability, Fluid Intelligence, and Crystallized Intelligence, respectively, at stratum II.

The subtests of the Dutch WISC-R [48]:

1. Information; the child has to verbally answer all kinds of general questions, some of which have several possible correct answers. This subtest measures General Information, which is a measure of Crystallized Intelligence at stratum II.
2. Picture Completion; the child has to find out which essential part of a picture is missing, within a given time. This subtest measures Closure Speed at stratum I, which makes it a measure of Broad Visual Perception at stratum II.
3. Similarities; the child has to find a similarity between two objects or concepts. There are several correct answers. This subtest is a measure of Induction at stratum I, which makes it a measure of Fluid Intelligence at stratum II.
4. Picture Arrangement; the child has to order a series of pictures in such a way that the pictures form a comprehensive story within a given time. This subtest is a measure of General Sequential Reasoning at stratum I, which makes it a measure of Fluid Intelligence at stratum II.
5. Arithmetic; the child has to solve arithmetic problems. These arithmetic problems are verbally presented: Four boys have 72 fish. They divided the fish, and everybody gets the same amount. How many fish does each boy get? This subtest is a measure of Crystallized Intelligence at stratum II.
6. Block Design; using blocks, the child has to replicate a pattern presented on a card. This subtest is a measure of Visualization at stratum I, which measures Broad Visual Perception at stratum II.
7. Vocabulary; the child has to give the meaning of a presented word. This subtest measures Lexical Knowledge at stratum I, which makes it a measure of Crystallized Intelligence at stratum II.
8. Object Assembly; the child has to put different pieces of cardboard together to copy a given figure within a given time. This subtest is a measure of Visualization at stratum I, which makes it a measure of Broad Visual Perception at stratum II.
9. Comprehension; the child has to answer different questions in which they have to give their insight and judgment about everyday-life issues. This subtest measures General Knowledge, which is a measure of Crystallized Intelligence at stratum II and is a measure of General Sequential Reasoning at stratum I, which is a measure of Fluid Intelligence at stratum II.
10. Coding; the child has to put a sign in a series of figures (code A) or under a series of numbers (code B). The sign belonging to the figure of number that was presented to the child earlier. This subtest is a measure of Visual Memory as stratum I, which falls under General Memory and Learning at stratum II.
11. Digit Span; the child has to repeat a series of numbers in the sequence presented to them auditorily (Forward Digit Span) or in reverse order starting with the last number they heard back to the first number (Backward Digit Span). This subtest is a measure of Memory Span at stratum I, which makes it a measure of General Memory and Learning at stratum II.
12. Mazes; the child has to trace the way out of a maze presented on paper with a pencil within a given time. The child is not allowed to enter a dead end. This subtest is a measure of Spatial Scanning at stratum I, which falls under Broad Visual Perception at stratum II.

2.9.3. *g* Loadings

The data in Table 1 are taken from the RAKIT manual [46] (p. 142, Table 9.4) and were computed using the same techniques as in the rest of the present study (see Section 2.4).

Table 1. Combined datasets of RAKIT and Dutch WISC-R: their subtests and *g* loadings.

Subtest		*g*
Dutch names	*English names*	
Rakit		
Figuur Herkennen	Figure Recognition	0.53
Exclusie	Exclusion	0.47
Geheugenspan	Memory Span	0.56
Woordbetekenis	Verbal Meaning	0.39
Doolhoven	Mazes	0.61
Analogieën	Analogies	0.63
Kwantiteit	Quantity	0.51
Schijven	Disks	0.49
Namen Leren	Learning Names	0.62
Verborgen figuren	Hidden Figures	0.20
Ideeënproduktie	Idea Production	0.25
Vertelplaat	Storytelling	0.58
WISC-R		
Informatie	Information	0.61
Onvolledige tekeningen	Picture Completion	0.52
Overeenkomsten	Similarities	0.57
Plaatjes ordenen	Picture Arrangement	0.70
Rekenen	Arithmetic	0.64
Blokpatronen	Block Design	0.56
Woordenschat	Vocabulary	0.43
Figuurleggen	Object Assembly	0.27
Begrijpen	Comprehension	0.35
Substitutie	Coding	0.44
Cijferreeksen	Digit Span	0.44
Doolhoven	Mazes	0.53

2.9.4. Computation of g Scores and "True" *g* Scores

The various *g* scores of all research participants were computed by summing the products of participant's *z* scores and the subtest's *g* values for all the subtests. "True" *g* scores were the *g* scores computed on the full set of 24 subtests. The other *g* scores were based on a minimum of five subtests and a maximum of 23 subtests.

2.9.5. Combinations of Subtests

Here we set five subtests as the minimum to create an artificial test battery. The maximum number of subtests for an artificial battery was set at 23. A basic artificial test battery with a well-balanced

combination of subtests was created by taking the 5 RAKIT subtests Closure (Broad Visual Perception), Exclusion (Fluid Intelligence), Memory Span (Memory), Verbal Meaning (Crystallized Intelligence), and Idea Production (Broad Retrieval Ability). These five subtests measure five of the broad dimensions of the Carroll [47] model that are most commonly represented in IQ batteries. They constitute subtests numbers 1, 2, 3, 4, and 11, respectively. Additional artificial test batteries were created by adding subtests to the basic artificial test battery. Nineteen artificial test batteries of six subtests were created by adding RAKIT subtests numbers 5, 6, 7, 8, 9, 10, and 12 and WISC-R subtests 1–12 to the basic battery. Eighteen artificial test batteries of seven subtests were created by adding RAKIT subtests numbers 5 and 6; 6 and 7; 7 and 8; 8 and 9; 9 and 10; and 10 and 12; RAKIT subtest 12 and WISC-R subtest 1; WISC-R subtests 1 and 2; etc., to the basic battery. Artificial batteries consisting of 8–23 subtests were created in a similar manner. *g* scores of these combinations were computed using the *g* loadings computed on the full sample.

2.9.6. Correlations of *g* Scores and "True" *g* Scores

Pearson correlations were computed between the *g* scores of the large number of artificial test batteries and the *g* score based on the total collection of subtests, which was taken as a measure of "true" *g*.

2.9.7. Scatter Plot

All the data points were entered into a scatter plot with number of subtests in an artificial battery on the x-axis and the correlation between the two sum scores on the y-axis. If the hypothesis about "true *g*" is correct, the scatter plot should show that the larger *N* becomes, the higher the value of the correlation, with an asymptotic function between *r* and *N* expected. The curve that gave the best fit to the expected asymptotic function was selected, and a logarithmic regression line was always tried first.

2.9.8. Computation of the Correction Value

The regression line of the scatter plot was used to estimate the correction values for the correction for imperfectly measuring the construct of *g*. First, the average number of subtests for every study in the meta-analytical database was computed. Second, the correction value was estimated by taking the cut-point of the regression line for the average number of subtests. In this way one correction factor applicable for the correction for deviation from perfect construct validity for the entire sample was obtained.

3. Results

A scatter plot of correlations of the *g* score of artificial test batteries with "true" *g* against number of subtests should reveal that the larger the number of subtests becomes, the higher the value of the correlation, with an asymptotic function between *r* and number of subtests expected, and this is what we found. Figure 3 shows the scatter plot of the correlations between two sum scores and number of subtests, and the logarithmic curve that fitted optimally for the data from the RAKIT and Dutch WISC-R combined.

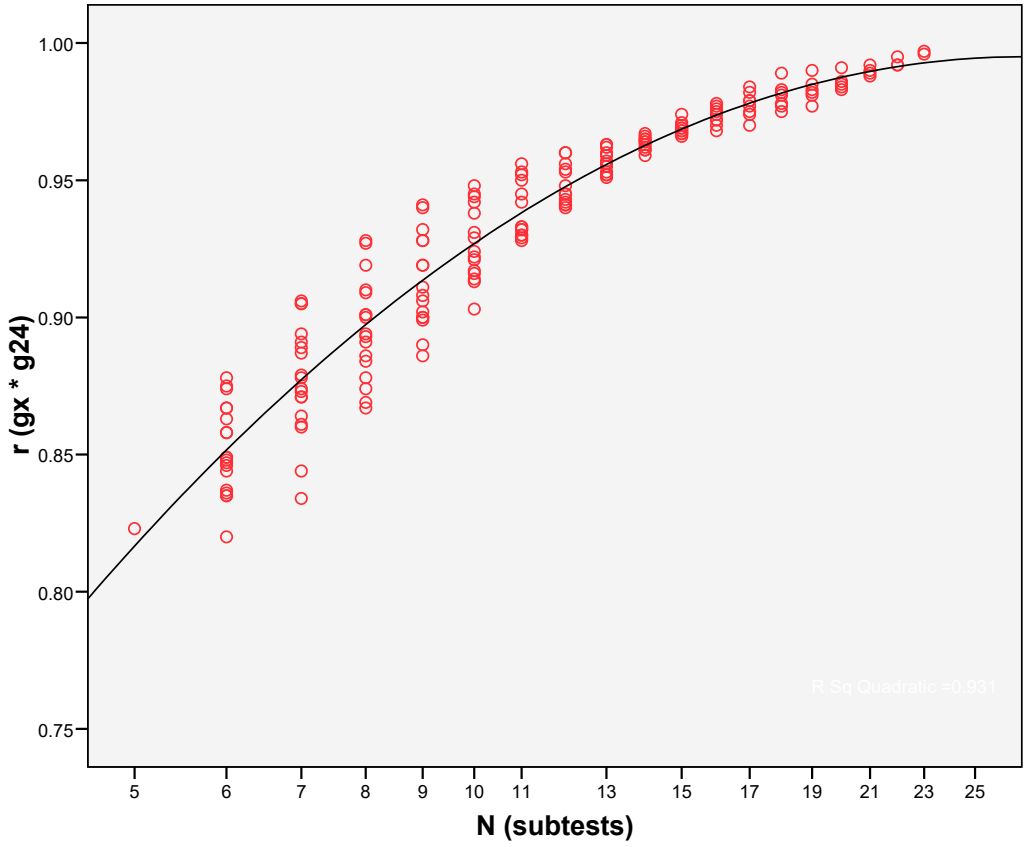


Figure 3. RAKIT and Dutch WISC-R combined: scatter plot of correlations of g score from artificial test battery with "true" *g* against number of subtests and regression line.

In the meta-analytical database, the number of subtests in the IQ batteries ranged from six to sixteen with the mean value is $m = 11.00$ ($SD = 2.96$). The correlations between *g* scores and "true"-*g* scores for the combined RAKIT and WISC-R dataset was estimated by taking the cut-point of the regression line based on eleven subtests, which results in $r = 0.94$. This means that a correction of 1/0.94 = 6% has to be applied to the value of rho-4. So, for instance, when the value of *rho*-4 = 0.50, the value of *rho*-5 becomes 0.53.

Table 2 gives an overview of the correlation between the Glass' *d* scores and *g* loadings for comparisons of Latin-American Hispanics and Whites. The Table contains data from a total of 13 studies, which yield a total of 14 correlations. The total number of White subjects is 230,974, while the total number of Hispanic subjects is 16,813. The combined harmonic mean of all the samples is 59,089. The first column reports the reference for the study, the second column reports the group (Hispanic, Mexican-American, or both), the third column reports the test battery used, and the fourth column reports the *r* (*g* × Glass' *d*) for Latin-American Hispanics. The fifth column reports the amount of subtests the test contains, the sixth column reports the size of the White samples, the seventh column reports the size of the Latin-American Hispanic samples, the eighth column reports the $N_{harmonic}$ for each comparison, the ninth column reports the *N* on which the computation of *g* loadings was based, the tenth column reports the reliability value of the *g* vector, the eleventh column reports the reliability of the Glass' *d* vector, the twelfth column reports the amount of restriction of range, expressed in *u*, and the thirteenth column reports the age of the subjects. It is clear that with one exception all the correlations are positive and many times Spearman's hypothesis is strongly supported.

Table 2. Overview of studies with correlations between the vector of *g* loadings and the vector of standardized Latin-American Hispanic/White differences on the subtests of an IQ battery.

Study	Group	Test Battery	*r*	$N_{subtests}$	N_{White}	$N_{Hispanic}$	N_{harmonic}	N_g	r_{gg}	r_{dd}	*u*	Mean Age (Range)
Carretta (1997) [49]	Hispanic	AFOQT	0.36	16	212,238	12,647	47,743	212,238	1.00	1.00	0.68	21 (18–27)
Reynolds, Willson, & Ramsey (1999) [50]	Mexican-American	WISC-R	0.77	12	2200	223	810	2200	0.98	0.87	0.77	10.12 (6–16)
Valencia & Rankin (1986) [51]	Mexican-American	K-ABC	0.70	13	100	100	200	1500	0.96	0.57	0.64	11 (10–12.5)
Hartmann, Kruuse, & Nyborg (2007) [14]	Hispanic	Various tests	0.71	16	3556	181	689	3556	0.98	0.84	0.99	19.9 (17–25)
Hartmann, Kruuse, & Nyborg (2007) [14]	Hispanic	ASVAB	0.74	10	6947	1704	5473	6947	0.99	1.00	0.59	19.6 (15–24)
Snyder [2] (1991) [52]	Hispanic	WISC-R	0.81	11	64	64	128	1800	0.96	0.45	0.68	10.5 [1] (6.5–14.5)
Dalliard (2013) [16]	Hispanic	DAS-II	0.70	13	864	432	1152	2952	0.98	1.00	0.73	(5–17)
Taylor & Richards (1991) [53]	Hispanic	WISC-R	0.72	10	1200	100	369	1200	0.96	0.70	0.56	8.3 (6–11)
Sandoval (1979) [54]	Mexican-American/Hispanic	WISC-R	0.55	10	351	349	700	1200	0.96	0.85	0.50	8 (5–11)
Kaufman, McLean & Kaufman (1995) [42]	Hispanic	KAIT	0.56	8	1535	138	502	1535	0.96	0.75	0.45	11–94
Dean[2] (1979) [55]	Mexican-American	WISC-R	0.75	10	60	60	120	2200	0.98	0.44	0.68	10
Naglieri, Rojahn, & Matto (2007) [56]	Hispanic	CAS	−0.47	12	1956	244	868	155	0.78	0.92	0.82	8.3 (5–17)
Kane (2007) [15]	Hispanic	UNIT	0.42	6	77	77	154	77	0.48	0.50	1.07	10.5
Flemmer & Roid (1997) [43]	Hispanic	Leiter-R	0.28	7	258	62	181	410	0.88	0.54	0.31	11–21

Note. *r* = correlation between Latin-American Hispanic and White differences; N_{subtests} is number of subtests in the intelligence battery; N_{White} = sample size for Whites; $N_{\text{Hispanics}}$ = sample size for Latin-American Hispanics; N_{harmonic} is computed using the formula $\frac{4}{\frac{1}{n1}+\frac{1}{n2}}$ where *n*1 and *n*2 are the number of participants in groups *n*1 and *n*2, respectively; N_g = sample size for *g* vector; r_{gg} is the reliability of the *g* vector; r_{dd} is the reliability of the Glass'*d* vector; *u* indicates the restriction of range. AFOQT = Air Force Officer Qualification Test, WISC-R = Wechsler Intelligence Test for Children-Revised, K-ABC = Kaufman Assessment Battery for Children, ASVAB = Armed Services Vocational Aptitude Battery, DAS-II = Differential Ability Scales-II, KAIT = Kaufman Adolescent and Adult Intelligence Test, CAS = Cognitive Assessment System, UNIT = Universal Nonverbal Intelligence Test, Leiter-R = Leiter International Performance Scale-Revised.[1] Estimated; [2] Referral sample; see text for explanation.

Table 3 presents the results of our meta-analysis based on a total of 14 data points. For detailed descriptions on the computation of the various variables we refer the reader to Hunter and Schmidt [44], which contains more of these descriptions than Schmidt and Hunter [29]. Table 3 reports the number of studies (*K*), the total sample size based on number of subtests in an IQ battery (total *N*), the uncorrected correlation between *d* and *g* (*r*), the standard deviation of *r* (SD_r), the correlation meta-analytically corrected for four statistical artifacts (*rho*-4), the standard deviation of *rho*-4 ($SD_{rho\text{-}4}$), the correlation meta-analytically corrected for five statistical artifacts (*rho*-5), the percentage of variance explained by four artifacts (%VE), and the 80% credibility interval (80% CI). The first four corrections for statistical artifacts are correction for sampling error, correction for unreliability of the *g* vector, correction for unreliability of the *d* vector, and correction for restriction of range, respectively; the fifth correction is the correction for imperfectly measuring the construct of *g*.

Table 3. Meta-analytical results for correlations between *g* loadings and Latin-American Hispanic/White differences.

Studies Included	*K*	Total *N*	*r*	SD_r	*rho*-4	$SD_{rho\text{-}4}$	*rho*-5	%VE	80% CI
All studies	14	154	0.55	0.333	0.75	0.368	0.80	24.3	0.27–1.22
All studies minus outlier	13	142	0.63	0.158	0.86	0	0.91	199.7	0.86–0.86

Note. *K* = number of correlations; total *N* = total sample size based on number of subtests in IQ battery; mean *r* = vector correlation weighted by number of subtests in IQ battery; SD_r = standard deviation of observed correlations; *rho*-4 = correlation meta-analytically corrected for four artifacts; $SD_{rho\text{-}4}$ = standard deviation of correlation meta-analytically corrected for four artifacts; *rho*-5 = correlation meta-analytically corrected for five artifacts; %VE = percentage of variance accounted for by four artifacts; 80% CI = 80% credibility interval, computed using $SD_{rho\text{-}4}$.

The overall analysis yields a mean observed correlation of 0.55, with a *rho*-5 of 0.80 and 24.3% of variance explained by four artifacts. However, there is a clear outlier in the data, namely the data point from Naglieri et al. [56] with a *r* (*g* × Glass' *d*) = −0.47. The mean of all the other 13 *r* (*g* × Glass' *d*)s is 0.63 (*SD* = 0.171), so the Naglieri et al. [56] data point is 6.37 *SD* below the mean and 4.39 *SD* below the lowest value in the database, namely the one from Flemmer and Roid [43]. We decided to leave the Naglieri et al. data point out, which strongly improved the outcomes in the sense that the amount of variance in the data points explained by statistical artifacts now became 199.7%. This is a case of second-order sampling error (see for a detailed explanation: [44]) and simply means that all the variance in the data points is explained by statistical artifacts.

4. Discussion

In this study we meta-analytically tested Spearman's hypothesis on Latin-American Hispanic samples. We initially found quite strong support for Spearman's hypothesis, but after removing both a clear outlier the mean *r* became much higher, which implies clear support for Spearman's hypothesis. Corrections for statistical artifacts substantially increased the value of the meta-analytical correlations. So, there was strong, meta-analytical support for Spearman's hypothesis.

Could MGCFA have been used to analyze the data points in the present meta-analysis? Dolan's [21] version of MGCFA requires access to the original datasets, or access to the means, *SD*s, and correlation matrices of the two or more groups being compared. We did not have access to the original datasets of the studies in the present meta-analysis, but we carefully checked these studies for the correlation matrices being reported; obviously, the means and *SD*s on at least one of the groups were being reported, otherwise we could not have used the method of correlated vectors. The correlation matrices were reported for a group of 100 Mexican-Americans and a group of 100 Whites in Valencia and Rankin [51], and for a group of 64 Latin-American Hispanics and 64 Whites in Snyder [52]; the other eleven studies did not report the correlation matrices. However, the samples in these two studies are simply too small to use MGCFA [57]. This means that none of the data points in the present meta-analyses could be reanalyzed using MGCFA.

Following te Nijenhuis and van den Hoek [35], we add a cautionary note concerning conditions that are not fulfilled in our study, which means that our conclusions are only conditionally valid. Measurement invariance is, strictly speaking, a necessary condition on a priori grounds. In the present study we could not use MGCFA, so we simply could not test for measurement invariance; moreover, we did not use Jensen's [28] procedure for testing Spearman's hypothesis, but a more simplified procedure [10], which does not allow testing for measurement invariance. This means that it is possible that some of the datasets could have shown lack of measurement invariance to a certain degree. We employed a trade-off where we collected a substantial number of studies that could only be analyzed using nonoptimal statistical techniques, but which allowed the use of meta-analysis, which is a powerful technique. We refer the interested reader to the meta-analysis on the effects of organizational development by Rodgers and Hunter [58], which describes, in detail, a trade-off leading to inclusion of many studies of lesser methodological quality allowing a huge meta-analysis.

The studies used in our meta-analysis describe the samples as either 'Hispanic' or 'Mexican-American'. It would be interesting to use more fine-grained ethnic distinctions. Is Spearman's hypothesis more strongly supported for Mexican-Americans hailing from parts of Mexico populated by people with an Amerindian background? How strongly does socioeconomic status of the group play a role? New studies reporting more detail on background would be welcome.

Cuban-Americans are Latin-American Hispanics but we did not find usable studies on them. Most likely Cuban-Americans have IQ scores that are comparable to those of Anglo-Americans, but it is also possible that they have slightly higher IQ scores. In the latter case, the scores of the Anglo-Americans will have to be subtracted from those of the higher-scoring Cuban-Americans for a proper test of Spearman's hypothesis, just as was done in the meta-analysis of Spearman's hypothesis tested on Jews, where European Jews had higher scores than non-Jewish Whites. This is in line with Jensen's logic of subtracting the IQ scores of the lower-scoring group from the IQ scores of the higher-scoring group. Whether Spearman's hypothesis will then be supported is an empirical question.

Te Nijenhuis, Willigers, Dragt, and van der Flier [59] show the outcomes of a number of meta-analyses employing the method of correlated vectors where sampling error, based on the number of test takers or the harmonic N, generally explained a very modest amount of variance in the data points in the meta-analysis. This is in stark contrast to studies on the predictive validity of IQ tests for job performance where sampling error often explained a large amount of variance between the data points [60]. However, in the present meta-analysis, we used the number of subtests in an IQ battery as a basis for computing sampling error, leading to a dramatic increase in the amount of variance explained in the data points. Reanalyses of these older meta-analyses will have to show to what degree the outcomes—the meta-analytical r and the percentage variance in the data points explained by statistical artifacts—change when using the number of subtests instead of the number of research participants as a basis to compute sampling error.

Te Nijenhuis, Bakhiet et al. [61] described how the method of correlated vectors has been applied to a large number of phenomena and all studies on cultural variables (such as Headstart gains, adoption gains, test–retest gains, and learning potential training gains) show a substantial to strong negative correlation with g loadings. In contrast, the majority of studies on biological–genetic variables (such as brain's glucose metabolic rate and heritability) show a strong positive correlation with g loadings. It is clear that the pattern in Latin-American Hispanic/White differences in intelligence is more similar to the pattern in biological–genetic variables than to the pattern in cultural variables. Correlational studies do not allow strong conclusions, so we conclude that the outcomes are suggestive that biological–genetic variables are more important than cultural variables in explaining Latin-American Hispanic/White differences.

With concern to the correction for imperfectly measuring the construct of g, previous studies [32,36] used a conservative value of 10% to limit the risk of overcorrection, but the new method for computing the correction for imperfectly measuring g suggests that the correction used before was too strong. In the present meta-analysis with quite a few IQ batteries with many subtests we estimated the value

of the correction factor to be 6%. However, additional studies using a large number of subtests are needed to see whether we get comparable outcomes.

The advantage of a theoretically-derived value of the unrestricted *SD* is that it can be employed in all psychometric meta-analyses of cognitive measures used to test Spearman's hypothesis. It makes the outcomes of all these psychometric meta-analyses more strongly comparable than when using empirically-derived values for the unrestricted *SD* for each specific field. Most likely, a battery of simple reaction time measures will have a substantially smaller *SD* than a battery of IQ tests. It would most likely also allow combining the meta-analyses for different instruments into one higher-order meta-analysis.

The higher the value of the unrestricted *SD*, the stronger the correction on the value of the observed correlation and the higher the value of *rho*. This means that the corrections for restriction of range based upon theoretically-derived values will lead to stronger meta-analytical support for Spearman's hypothesis then when using empirically-derived values of *SD*. However, it is important to limit the theoretical risk of overcorrection, for instance by checking whether the value of *rho*-5 (the value of the correlation after having applied all the five corrections for statistical artifacts) does not become substantially larger than 1. A value of, say, *rho*-5 = 1.05 would not be a big problem, reflecting simply a small amount of overcorrection that could be tried to remedy by basing the corrections on more observations. However, a value of, say, *rho*-5 = 1.2 would suggest there is a fundamental flaw somewhere in the corrections. Jensen [10] states that the correction for restriction of range is the strongest of the various corrections for statistical artifacts: it leads to the biggest changes in the observed correlation. So, the value of unrestricted *SD* chosen generally has the most powerful influence on the value of *rho*-5. Therefore, arriving at a value of *rho*-5 = 1.2 would first of all throw doubt on the solidness on the correction for restriction of range, and it would lead one to question using the theoretically-derived values of unrestricted *SD*. This value of unrestricted *SD* should also be tested for overcorrection in other meta-analyses testing Spearman's hypothesis using cognitive measures: reaction time measures, SJTs, school achievement measures, etc. If there are overcorrections with most cognitive measures, then the solidness of the chosen correction for restriction can be disputed.

A reviewer commented upon our choice of *SD*s when computing effect sizes and suggested using pooled *SD*s. In most tests of Spearman's hypothesis, a White group is compared to a non-White group, and very often the White group is much larger than the non-White group. The Carretta (1997) [49] study is a nice example, because the White group is more than 16 times as large as the Hispanic group. In such a case, using the *N*-weighted average *SD* would yield virtually identical conclusions to using the White *SD*, as the White *SD* would very strongly influence the *N*-weighted average *SD*. Of the 14 data points in our study, nine showed much larger to very much larger *N*s for the White comparison group and only five showed comparable sample sizes. Using the *N*-weighted average *SD* will have at best a very small effect on the weighted mean *r* of the nine studies. As the five studies with comparable sample sizes are only a third of the meta-analytical database, we are quite sure that using the procedure suggested by the reviewer will lead to highly comparable outcomes to what we report now. Please note that we always try to use the best quality *SD*s, namely those from nationally representative samples. Sample *SD*s are approximations of populations *SD*s, with the values of *SD*s of small samples showing a lot of sampling error. However, the *SD*s from carefully collected nationally representative samples were very close to the population value.

There is a limitation to our study, which is inherent to virtually all uses of meta-analysis. We strongly believe that the only amount of scientific information that can be taken seriously is the amount of information contained in a good-sized meta-analysis [62]. The overwhelming majority of meta-analyses is based on the information reported in published studies; access to the original studies is generally impossible, especially when studies are relatively old. We simply had no access to the original data in all the studies included in our meta-analysis, which limited the amount of data we could use and the number of statistical analyses we could carry out. Luckily, the study by Hartmann, Kruuse, and Nyborg (2007) [14], which supplies two data points to our meta-analysis, was carried

out on the original datasets and it was tested whether Jensen's requirement for similar *g* loadings for Whites and Hispanics were met. The authors report a strong similarity in *g* loadings, with all congruence coefficients > 0.98.

We conclude that, as we hypothesized, Spearman's hypothesis was strongly supported for Latin-American Hispanic/White differences. From this we can conclude that the differences between Whites and Latin-American Hispanics are primarily caused by differences in general intelligence. Spearman's hypothesis has now not only been supported for Blacks, Jews, and Amerindians, but also for Latin-American Hispanics. It would be of interest to find groups for which Spearman's hypothesis is not supported.

Author Contributions: Conceptualization, J.t.N.; Methodology, J.t.N.; Formal Analysis, J.t.N, M.v.d.H., and J.D.; Investigation, J.t.N, M.v.d.H., and J.D.; Resources, J.t.N, M.v.d.H., and J.D.; Data Curation, J.t.N, M.v.d.H., and J.D.; Writing—Original Draft Preparation, J.t.N, M.v.d.H., and J.D.; Writing—Review & Editing, J.t.N.; Visualization, J.t.N and M.v.d.H., J.D.; Supervision, J.t.N.; Project Administration, J.t.N, M.v.d.H., and J.D.

Funding: This research received no external funding.

Acknowledgments: Thanks to Nico Bleichrodt and Wilma Resing for giving access to the dataset on the Dutch WISC-R and the RAKIT. Thanks to Hebe van der Hoeff for helping to clean the data for the computation of the reliability of the *g* vectors.

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

References

1. U.S. Department of Commerce, Economics and Statistics Administration, U.S. Census Bureau. The Hispanic Population: 2010: 2010 Census Briefs (Publication No. C2010BR-04). 2011. Available online: http://www.census.gov/prod/cen2010/briefs/c2010br-04.pdf (accessed on 15 January 2019).
2. U.S. Census Bureau, National Advisory Committee on Racial, Ethnic and Other Populations. 2020 Census: Race and Hispanic Origin Research Working Group: Final Report. 2014. Available online: https://www2.census.gov/cac/nac/reports/2014-06-10_RHO_wg-report.pdf (accessed on 15 January 2019).
3. CIA Factbook. The World Factbook. 2015. Available online: https://www.cia.gov/library/publications/the-world-factbook/geos/us.html (accessed on 15 January 2019).
4. Lynn, R.; Vanhanen, T. *IQ and the Wealth of Nations*; Praeger: London, UK, 2002.
5. Kena, G.; Aud, S.; Johnson, F.; Wang, X.; Zhang, J.; Rathbun, A.; Wilkinson-Flicker, S.; Kristapovich, P. *The Condition of Education 2014 (NCES 2014-083)*; U.S. Department of Education, National Center for Education Statistics: Washington, DC, USA, 2014. Available online: http://nces.ed.gov/pubsearch (accessed on 15 January 2019).
6. Ryan, C.L.; Siebens, J. *Educational Attainment in the United States: 2009*; US Census Bureau: Washington, DC, USA, 2012.
7. Brown, A.; Patten, E. *Statistical Portrait of Hispanics in the United States, 2012*; Pew Hispanic Center: Washington, DC, USA, 2014.
8. Lynn, R. Racial and ethnic differences in intelligence in the United States on the Differential Ability Scale. *Personal. Individ. Differ.* **1996**, *20*, 271–273. [CrossRef]
9. Roth, P.L.; Bevier, C.A.; Bobko, P.; Switzer, F.S.; Tyler, P. Ethnic group differences in cognitive ability in employment and educational settings: A meta-analysis. *Pers. Psychol.* **2001**, *54*, 297–330. [CrossRef]
10. Jensen, A.R. *The g Factor: The Science of Mental Ability*; Praeger: Westport, CT, USA, 1998.
11. Rushton, J.P.; Jensen, A.R. Thirty years of research on race differences in cognitive ability. *Psychol. Publ. Pol. Law* **2005**, *11*, 235–294. [CrossRef]
12. te Nijenhuis, J.; van den Hoek, M.; Armstrong, E. Spearman's hypothesis and Amerindians: A meta-analysis. *Intelligence* **2015**, *50*, 87–92. [CrossRef]
13. 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]
14. Hartmann, P.; Kruuse, N.H.S.; Nyborg, H. Testing the cross-racial generality of Spearman's hypothesis in two samples. *Intelligence* **2007**, *35*, 47–57. [CrossRef]

15. Kane, H. Race differences on the UNIT: Evidence from multi-sample confirmatory analysis. *Mank. Q.* **2008**, *48*, 283–298.
16. Dalliard. Spearman's Hypothesis and Racial Differences on the DAS-II. Humanvarieties.com. 2013. Available online: http://humanvarieties.org/2013/12/08/spearmans-hypothesis-and-racial-differences-on-the-das-ii/ (accessed on 15 January 2019).
17. Ganzach, Y. Another look at the Spearman's hypothesis and relationship between Digit Span and General Mental Ability. *Learn. Individ. Differ.* **2016**, *45*, 128–132. [CrossRef]
18. Ganzach, Y. On general mental ability, digit span and Spearman's hypothesis. *Learn. Individ. Differ.* **2016**, *45*, 135–136. [CrossRef]
19. Jensen, A.R.; Figueroa, R.A. Forward and backward digit span interaction with race and IQ. Predicitons form Jensen's theory. *J. Educ. Psychol.* **1975**, *67*, 882–893. [CrossRef]
20. Ashton, M.C.; Lee, K. Problems with the method of correlated vectors. *Intelligence* **2005**, *33*, 431–444. [CrossRef]
21. Dolan, C.V. Investigating Spearman's hypothesis by means of multi-group confirmatory factor analysis. *Multivar. Behav. Res.* **2000**, *35*, 21–50. [CrossRef] [PubMed]
22. Hunt, E. *Human Intelligence*; Cambridge University Press: Cambridge, UK, 2011.
23. 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]
24. te Nijenhuis, J.; van den Hoek, M. Spearman's hypothesis tested on Black adults: A meta-analysis. *J. Intell.* **2016**, *4*, 6. [CrossRef]
25. Te Nijenhuis, J.; Choi, Y.Y.; van den Hoek, M.; Valueva, E.; Lee, K.H. Spearman's hypothesis tested comparing Korean young adults with various other groups of young adults on the items of the Advanced Progressive Matrices. *J. Biosoc. Sci.* **2019**, in press.
26. Wicherts, J.M. Ignoring psychometric problems in the study of group differences in cognitive test performance. *J. Biosoc. Sci.* **2018**, *50*, 868–869. [CrossRef]
27. te Nijenhuis, J.; van den Hoek, M. Analysing group differences in intelligence using the psychometric meta-analytic-method of correlated vectors hybrid model: A reply to Wicherts (2018) attacking a strawman. *J. Biosoc. Sci.* **2018**, *50*, 870–871. [CrossRef]
28. Jensen, A.R. Spearman's hypothesis tested with chronometric information-processing tasks. *Intelligence* **1993**, *17*, 47–77. [CrossRef]
29. Schmidt, F.L.; Hunter, J.E. *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*, 3rd ed.; Sage: Thousand Oaks, CA, USA, 2015.
30. Schmidt, F.L.; Le, H. *Software for the Hunter-Schmidt Meta-Analysis Methods*; University of Iowa, Department of Management and Organization: Iowa City, IQ, USA, 2004.
31. Jensen, A.R. The nature of the black-white difference on various psychometric tests. Spearman's hypothesis. *Behav. Brain Sci.* **1985**, *8*, 193–263. [CrossRef]
32. te Nijenhuis, J.; van Vianen, A.E.M.; van der Flier, H. Score gains on g-loaded tests: No g. *Intelligence* **2007**, *35*, 283–300. [CrossRef]
33. te Nijenhuis, J.; Jongeneel-Grimen, B.; Kirkegaard, E.O. Are Headstart gains on the g factor? A meta-analysis. *Intelligence* **2014**, *46*, 209–215. [CrossRef]
34. te Nijenhuis, J.; Jongeneel-Grimen, B.; Armstrong, E.L. Are adoption gains on the g factor? A meta-analysis. *Personal. Individ. Differ.* **2015**, *73*, 56–60. [CrossRef]
35. te Nijenhuis, J.; van den Hoek, M.; Williger, D. Testing Spearman's hypothesis with alternative intelligence tests: A meta-analysis. *Mank. Q.* **2017**, *57*, 687–705.
36. te Nijenhuis, J.; van der Flier, H. Is the Flynn effect on g?: A meta-analysis. *Intelligence* **2013**, *41*, 802–807.
37. 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**, *44*, 1–10. [CrossRef]
38. Sternberg, R.J.; The Rainbow Project Collaborators. The Rainbow Project: Enhancing the SAT through assessments of analytical, practical, and creative skills. *Intelligence* **2006**, *34*, 321–350. [CrossRef]
39. Lynn, R. *Race Differences in Intelligence: An Evolutionary Analysis*; Washington Summit Books: Atlanta, GA, USA, 2006.
40. Glass, G.V.; McGaw, B.; Smith, M.L. *Meta-Analysis in Social Research*; Sage: London, UK, 1981.

41. Grissom, R.J.; Kim, J.J. *Effect Sizes for Research: Univariate and Multivariate Applications*, 2nd ed.; Routledge: New York, NY, USA, 2012.
42. Kaufman, A.S.; McLean, J.E.; Kaufman, J.C. The fluid and crystallized abilities of white, black, and Hispanic adolescents and adults, both with and without an education covariate. *J. Clin. Psychol.* **1995**, *51*, 636–647. [CrossRef]
43. Flemmer, D.D.; Roid, G.H. Nonverbal intellectual assessment of Hispanic and speech-impaired adolescents. *Psychol. Rep.* **1997**, *80*, 1115–1122. [CrossRef]
44. 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.
45. Flynn, J.R. *Intelligence and Human Progress: The Story of What was Hidden in our Genes*; Academic Press: Oxford, UK, 2013.
46. Bleichrodt, N.; Resing, W.C.M.; Drenth, P.J.D.; Zaal, J.N. *Intelligentiemeting Bij Kinderen [The Measurement of Children's Intelligence]*; Swets: Lisse, The Netherlands, 1987.
47. Carroll, J.B. *Human Cognitive Abilities*; Cambridge University Press: Cambridge, UK, 1993.
48. van Haasen, P.P.; de Bruyn, E.E.J.; Pijl, Y.J.; Poortinga, Y.H.; Lutje-Spelberg, H.C.; Vandersteene, G.; Coetsier, P.; Spoelders-Claes, R.; Stinissen, J. *Wechsler Intelligence Scale for Children-Revised, Dutch Version*; Swets: Lisse, The Netherlands, 1986.
49. Carretta, T.R. Group differences on US Air Force pilot selection tests. *Int. J. Sel. Assess.* **1997**, *5*, 115–127. [CrossRef]
50. Reynolds, C.R.; Willson, V.L.; Ramsey, M. Intellectual differences among Mexican Americans, Papagos and Whites, independent of g. *Personal. Individ. Differ.* **1999**, *27*, 1181–1187. [CrossRef]
51. Valencia, R.R.; Rankin, R.J. Factor Analysis of the K–ABC for groups of Anglo and Mexican American children. *J. Educ. Meas.* **1986**, *23*, 209–219. [CrossRef]
52. Snyder, B.J. WISC-R Performance Patterns of Referred Anglo, Hispanic, and American Indian Children. Ph.D. Thesis, University of Arizona, Tucson, AZ, USA, 1991.
53. Taylor, R.L.; Richards, S.B. Patterns of intellectual differences of Black, Hispanic, and White children. *Psychol. Sch.* **1991**, *28*, 5–9. [CrossRef]
54. Sandoval, J. The WISC-R and internal evidence of test bias with minority groups. *J. Consult. Clin. Psychol.* **1979**, *47*, 919–927. [CrossRef]
55. Dean, R.S. Distinguishing patterns for Mexican-American children on the WISC-R. *J. Clin. Psychol.* **1979**, *35*, 790–794. [CrossRef]
56. Naglieri, J.A.; Rojahn, J.; Matto, H.C. Hispanic and non-Hispanic children's performance on PASS cognitive processes and achievement. *Intelligence* **2007**, *35*, 568–579. [CrossRef]
57. Kline, R.B. *Principles and Practice of Structural Equation Modeling*, 3rd ed.; Guilford: New York, NY, USA, 2011.
58. Rodgers, R.; Hunter, J.E. The methodological war of the "Hardheads" versus the "softheads". *J. Appl. Behav. Sci.* **1996**, *32*, 189–208. [CrossRef]
59. te Nijenhuis, J.; Willigers, D.; Dragt, J.; van der Flier, H. The effects of language bias and cultural bias estimated using the method of correlated vectors on a large database of IQ comparisons between native Dutch and ethnic minority immigrants from non-Western countries. *Intelligence* **2016**, *54*, 117–135. [CrossRef]
60. Schmidt, F.L.; Hunter, J.E. The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychol. Bull.* **1998**, *124*, 262–274. [CrossRef]
61. 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]
62. Schmidt, F.L. What do data really mean?: Research findings, meta-analysis, and cumulative knowledge in psychology. *Am. Psychol.* **1992**, *47*, 1173–1181. [CrossRef]