Causes of group differences studied with the method of correlated vectors: A psychometric meta-analysis of Spearman's hypothesis

Joep Dragt
0263699

Master thesis
University of Amsterdam
Work and Organizational Psychology
Supervisor: Dr. Jan te Nijenhuis
March 2010

Acknowledgement

I would like to thank Dr. Jan te Nijenhuis and Dennys Franssen M.Sc. for their enthusiastic support and helpful guidance on this project. Furthermore, I would like to thank both Dr. Helmuth Nyborg, Dr. Nico Bleichrodt, and Dr. Wilma Resing for providing the large datasets, which were used to estimate the values for the correction for imperfectly measuring *g*. I would also like to thank the following researchers helped for additional information or studies on Spearman's hypothesis: Dr. Helmuth Nyborg, Dr. Christopher Brand, Dr. Fons van de Vijver, Dr. Peter Hartmann, Dr. Joe Fagan, Dr. Michael McDaniel, Dr. Cecil Reynolds, Dr. Richard Lynn, Dr. Phil Rushton, Dr. Michelle Helms-Lorenz, and Dr. Bryan Pesta. Finally, I want to thank Dr. John DeFries and Dr. Craig Nagoshi for their quick response to our request for additional information on their 1984 article in *Personality and Individual Differences*.

Table of contents

Abstract

Many studies have been conducted on differences in mean intelligence test scores between ethnic groups. Often cultural factors are used to explain these differences. However, Spearman's hypothesis states that the group differences on the subtests of IQ batteries can best be explained in terms of differences in the complexity of the tasks, that is, the demands they make on the general factor of mental ability, the *g* factor. In this paper a psychometric meta-analysis of Spearman's hypothesis was carried out. We predicted a strong positive correlation between ethnic group differences on the subtests of an IQ battery on the one hand, and the vector of *g* loadings on the other hand. Furthermore, subtests with a substantial verbal component measure to an undesirable extent proficiency in the language of the test taken and underestimate the level of *g* of the tested nonnative speakers. The questions which of the different subtests of the IQ batteries are language-biased and how large is the underestimation of IQ due to these language-biased subtests were addressed. Also, a separate study was carried out to empirically estimate the values for the correction for deviation from perfect construct validity in the psychometric meta-analysis.

Confirming our predictions, the psychometric meta-analysis of IQ batteries showed a strong relation between ethnic group differences and complexity: a true correlations of .91, based on a very large total *N*. Group identity was tested as a moderator, but no evidence was found: all different ethnic groups showed similar or in some cases virtually identical results. Our study on language bias showed that language-biased subtests underestimate IQ for minority groups, but virtually all effects are small: a mean underestimation was found of only 2.71 IQ points for all studies. Also, our study on the correction value for imperfectly measuring the construct of *g* estimated it to be 7.5 %, which is substantially smaller than the value used in previous research.

Spearman’s hypothesis can now be considered to be an empirical fact. Mean differences in intelligence between ethnic groups can be largely explained by the complexity of the subtests in an IQ battery. So, the present study shows clearly that there is simply no support for cultural bias as an explanation of these ethnic group differences. Apart from subtests with a strong language component, IQ batteries appear to be excellent measures of intelligence for all groups studied in our meta-analysis.

Differences in IQ test score are related to variables with immense practical consequences in today's technological society. In first world countries, everyday life itself has become something on an IQ test. The substantial correlation of IQ with many educational, economic, and social criteria is well established. As a result, cognitive tests are widely used for selection and placement in organizations, and increasingly also in educational settings. Largely for this reason, there has been a long-standing interest in the mean differences in IQ test score between various populations. By far the most extensively researched is that between the two largest populations in the United States: persons of European ancestry who are socially identified as ‘‘White’’ and persons of some African ancestry who are socially identified as ‘‘Black’’ or African-American (Jensen, 1998). Also, extensive research has been carried out on IQ differences between the Dutch and immigrants from non-Western countries (Evers, te Nijenhuis, & van der Flier, 2005).

The approximately normal distribution of IQ, as measured by nationally standardized tests, shows that, on average, the American Black population scores below the White population by about 1.2 standard deviations, equivalent to about eighteen IQ points. This statistical mean difference has changed little if at all over the past eighty years for which IQ data have been available (Jensen, 1980, 1998). Dutch-first-generation immigrants differences are about the same size as the US Black/White differences, but become substantially smaller for the second generation of immigrants (te Nijenhuis, de Jong, Evers, & van der Flier, 2004). The average difference, of course, is relatively small compared to the range of variation within either population and, in fact, is not much greater than the average difference between full siblings reared together in the same family.

The most visible educational, economic, and social consequences of the group differences in IQ arise largely from two effects: (1) the statistical characteristics of the normal curve, and (2) the minimum probable threshold of the level of ability needed for certain socially valued attainments, especially in education and in the job market. When two normal distributions of IQ have different means, although the curves largely overlap one another, a given cut-score on the IQ scale can make a very large difference between the proportions of the lower-scoring group and the higher-scoring group that fall below (or above) the cut-score. The further the distance of the cut-score from the mean of the higher-scoring group, the larger is the group difference between the proportion of each group that falls above (or below) the cut-score. Cut-scores on the IQ scale that fall at critical thresholds (mental retardation, passing grades in regular classes, high school graduation, college admission, college degree, high-level occupation, and the like) therefore result in significant disparities between the proportions of the higher- and the lower-scoring groups that reach the threshold for different social and occupational categories. Therefore it is important to investigate the nature and causes of these groups disparities (Jensen, 1998).

Extensive research on test bias has shown that no fraction of the Black/White (B/W) IQ differences, at least in the United States is attributable to any cultural bias in the test instruments, as such. Nor is the magnitude of the differences a function of the formal characteristics of the tests, such as verbal, nonverbal, individual versus group administration, culture-loaded, or culture-reduced. For all their legitimate, practical, and typical uses, present-day psychometric tests of mental ability have the same reliability and validity for native, English-speaking Blacks (as well as American-born, English speaking Hispanics and Asians) as they have for Whites (Jensen, 1980). Extensive research on test bias against non-Western immigrants in the Netherlands only yields convincing proof of language bias (Evers et al., 2005).

General Intelligence (g)

A well-established empirical finding—the manifold of positive correlations among measures of various mental abilities—is putative evidence of a general factor in all of the measured abilities. The method of factor analysis makes it possible to determine the degree to which each of the variables is correlated (or loaded) with the factor that is common to all the variables in the analysis. Spearman termed this *g* to represent a general factor that is manifested in individual differences on all mental tests, regardless of content (Jensen, 1998, p. 18). Spearman's *g* is best understood as a measure of cognitive complexity (Gottfredson, 1997), and 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 (Jensen & Weng, 1994). Thus, tests demanding higher cognitive complexity are high on *g* (have high *g* loadings), and tests demanding lower cognitive complexity are low on *g* (have low *g* loadings).

Hierarchical Intelligence Model

Jensen (1998) hypothesized that scores on IQ batteries are best described by hierarchical intelligence models, such as Carroll's (1993) three-stratum hierarchical factor model of cognitive abilities. At the highest level of the hierarchy (stratum III) is general intelligence or *g*. One level lower (stratum II) is occupied by the broad abilities of Fluid Intelligence, Crystallized Intelligence, General Memory and Learning, Broad Visual Perception, Broad Auditory Perception, Broad Retrieval Ability, and Broad Cognitive Speediness or General Psychomotor Speed. One level lower still (stratum I) comprises the narrow abilities, including Sequential Reasoning, Quantitative Reasoning, Verbal Abilities, Memory Span, Visualization, and Perceptual Speed. The lowest level of the hierarchy consists of large numbers of specific tests and subtests. Some tests, despite seemingly very different formats, have empirically demonstrated to cluster into one narrow ability (Carroll, 1993).

Method of Correlated Vectors (MCV)

The MCV is a means of identifying variables that are associated with Spearman's *g*, the general factor of mental ability. This method involves calculating the correlation between: (a) the column vector of the *g* factor loadings of the subtests of an intelligence test or similar battery, and (b) the column vector of the relation of each of those same subtests with the variable in question. When the latter variable is dichotomous, the relations are usually calculated in terms of an effect size. When the latter variable is continuous (or nearly so), the relations are usually calculated in terms of a correlation coefficient (Ashton & Lee, 2005).

Spearman's hypothesis

The magnitude of the mean Black/White difference varies considerably across tests as a function of the homogeneity of the test items. However, this variation between tests in the size of the standardized mean Black/White difference is not explainable in terms of test bias or in terms of differences in types of item content or other formal or superficial characteristics of the tests. Charles Spearman (1927) suggested that the different relative magnitudes of the Black/White differences on various tests are a function of each test's *g* loading, that is, the demands that solving the item correctly makes on general mental ability. This hypothesis (now called "Spearman's hypothesis") has since been tested in numerous studies based on large, representative samples of the American Black and White populations and has been strongly confirmed for them. The degree to which a particular test is *g*-loaded predicts the magnitude of the standardized mean Black/White difference on that test better than does any other psychometric factor yet identified. This implies that the Black/White difference consists mainly of difference in *g* (Jensen, 1998).

Spearman's hypothesis has also been studied using other methods than intelligence tests. First, there are elementary cognitive tasks (ECTs) which measure the time it takes a person to process information presented in tasks that are so simple that virtually all persons in the study sample are able to perform them correctly in only one or two seconds. The chronometric variables derived from such ECTs vary in their *g* loadings and show significant Black/White differences. The extent to which the different ECT variables are *g*-loaded predicts the relative magnitudes of the standardized mean B/W differences on the chronometric variables derived from the ECTs. Spearman's hypothesis is thus confirmed even for tasks that do not call upon previously acquired knowledge or skills and that scarcely resemble conventional psychometric tests (Jensen, 1993).

Second, Spearman's hypothesis has also been studied using Situational Judgment Tests and Assessment Center exercises, which are widely used in industrial- and organizational psychology. SJTs assess an applicant's judgment regarding situations encountered in the workplace (McDaniel, Hartman, Whetzel, & Grubb, 2007; McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001). Assessment Centers are a collection of predictors used primarily for the selection, promotion and/or development of higher-level managerial jobs using simulations of essential parts of a job (Cascio, 1991; Heneman & Heneman, 1994). Whetzel, McDaniel, and Nguyen (2008) tested

Spearman’s hypothesis on Situational Judgment Test performance and the cognitive loading of a SJT was defined by the extent to which the test correlated with cognitive ability. Whetzel et al.'s (2008) meta-analysis shows that mean race differences between Black, Hispanic, Asian, and White examinees in SJT performance are largely explained by the cognitive loading of the SJT such that the larger the cognitive load, the larger the mean race differences. For example, the correlation between cognitive complexity of the SJT and Black/White differences was .77. Goldstein, Yusko, Braverman, Smith, and Chung (1998) tested whether the cognitive complexity of an Assessment Center exercise was a predictor of subgroup differences. Goldstein et al. (1998) found that when cognitive complexity was removed Black/White differences were reduced to nonsignificance for all of the Assessment Center exercises in this study. This outcome gives strong support to the hypothesis that subgroup differences are a function of the cognitive complexity of the different exercises. Furthermore, Goldstein, Yusko, and Nicolopoulos (2001) explored Black/White differences in managerial compentencies. These researchers concluded that significant subgroup differences emerged for a majority of the more cognitively-loaded competencies (e.g., judgment), whereas nonsignificant differences were associated with the majority of the less cognitively-loaded competencies (e.g., human relations). Finally, Roth, Bevier, Bobko, Switzer, and Tyler’s (2001) meta-analysis of ethnic group differences in cognitive ability in employment and educational settings show differences that are largest on the most *g*-loaded measures.

Although Spearman's hypothesis was originally only applied to the mean Black/White difference, the hypothesis of mean group differences as a function of cognitive complexity has also been tested for Hispanic, Native-American, American-Asian, and Native-Hawaiian groups. The outcomes of these studies are similar to those for Black samples. Moreover, the hypothesis that group differences are a function of subtests' *g* loadedness has also been tested in the Netherlands, South Africa, Zimbabwe, Asia, and Serbia. In the large majority of these studies, group differences are strongly predicted by *g* loadings.

In the Netherlands comparisons were made between native Dutch individuals and immigrants from Surinam, the Netherlands Antilles, Morocco, and Turkey. Support for Spearman's hypothesis was found in samples of adults (te Nijenhuis & van der Flier, 1997; 2005), secondary school children (te Nijenhuis, Evers, & Mur, 2000) and young children (te Nijenhuis, Tolboom, Resing, & Bleichrodt, 2004). In South Africa comparisons were made of Blacks, Whites, and Indians (Lynn & Owen, 1994). Rushton and Jensen (2003) compared a US White sample with a Zimbabwean Black sample. Furthermore, in a series of studies by Lynn, Eysenck, and Jensen scores on various reaction time measures were compared for White Western samples, Black African samples, and Asian samples (Ja-Song & Lynn, 1992; Jensen, 1993; Jensen & Whang, 1993, 1994; Lynn, Chan, & Eysenck, 1991; Lynn & Holmshaw, 1990; Lynn & Shigehisa, 1991; Vernon &

Jensen, 1984).

Further, Rushton tested Spearman's hypothesis in a series of studies at the item level using Raven's Progressive Matrices in Africa and in Serbia (Rushton, 2002; Rushton, Čvorović, & Bons, 2007; Rushton & Skuy, 2000; Rushton, Skuy, & Fridjhon, 2002; Rushton, Skuy, & Fridjhon, 2003). Rushton and Skuy (2000) found that African/White differences were greater on those items of the SPM with the highest item-total correlations, which are the best measures of general factor of intelligence. More recently, Rushton et al. (2007) found that differences in intelligence between the Roma (Gypsy) community of Serbia and majority Serbians were found to be most pronounced on the most *g*-loaded items of the Raven's.

Beside IQ batteries, reaction time measures, and Raven's Progressive Matrices, Spearman's hypothesis has also been tested in educational settings in the Netherlands (te Nijenhuis, Evers, & Mur 2000; te Nijenhuis, Tolboom, Resing, & Bleichrodt, 2004; te Nijenhuis & van der Flier, 2005). For example, te Nijenhuis, Evers and Mur (2000) demonstrated that group differences between minority and majority group members on school criteria were predominantly accounted for by the *g*-loadedness of these criteria.

In sum, Spearman's hypothesis was confirmed in the large majority of comparisons of various groups and for all assessment instruments studied.

General Method

Psychometric meta-analysis

Psychometric meta-analysis (Hunter & Schmidt, 1990) estimates what the results of studies would have been if all studies had been conducted without methodological limitations or flaws. The results of perfectly conducted studies would allow a clearer view of underlying construct-level relationships (Schmidt & Hunter, 1999). The goal of the present psychometric meta-analyses is to provide a reliable estimate of the true correlation between group differences and the magnitude of *g* loadings in intelligence batteries.

In general, *g* loadings were computed by submitting a correlation matrix to a principal-axis factor analysis and using the loadings of the subtests on the first unrotated factor. In some cases *g* loadings were taken from studies where other procedures were followed; these procedures have been shown empirically to lead to highly comparable results. Finally, Pearson correlations between group differences and *g* loadings were computed.

There has been a discussion whether one should use Pearson *r* or Spearman's rho when applying the method of correlated vectors. The answer depends on whether one assumes an interval or an ordinal measurement level for IQ scores. Ranking of IQ scores can be seen as a way of categorizing intelligence levels on an ordinal scale. For instance, an IQ score of 150 indicates a higher level of intelligence compared to an IQ score of 75. However, the inference that an IQ score of 150 indicates a doubling in level of intelligence compared to an IQ score of 75 cannot be drawn.

In order to obtain our results, mean IQ scores were used to calculate the score differences between groups (*d*). Score differences have the characteristics of an interval scale: arithmetical operations can be conducted, and the effects (*d*) have values ranging from negative to positive. Thus, the choice for Pearson *r* or Spearman's rho depends on whether the underlying construct on which calculations are carried out are more important or the calculations themselves. Colom, Juan-Espinosa, Abad, and Garcia (2000) consider both Pearson *r* and Spearman's rho as suitable measures of the degree of relationship between two vectors. We decided to use Pearson *r* following earlier conducted meta-analyses using Pearson *r* in the method of correlated vectors (te Nijenhuis, van Vianen, & van der Flier, 2007; te Nijenhuis & Jongeneel-Grimen, 2007). This has the advantage that the results of the present studies can be compared directly against those of the earlier studies.

Corrections for Artifacts

Psychometric meta-analytical techniques (Hunter & Schmidt, 1990, 2004) were applied using the software package developed by Schmidt and Le (2004). Psychometric meta-analysis is based on the principle that there are artifacts in every dataset and that most of these artifacts can be corrected. In the present meta-analyses we corrected for five artifacts identified by Hunter and Schmidt (1990) that alter the value of outcome measures. These are: (1) sampling error, (2) reliability of the vector of *g* loadings, (3) reliability of the vector of a specific variable of theoretical interest (4) restriction of range of *g* loadings, and (5) deviation from perfect construct validity. In the present exploratory studies, using bare-bones meta-analytical techniques, we corrected for only one artifact, namely sampling error.

Study 1: Effect of Language Bias in Subtests

When comparing test scores of people who lack a desirable level of proficiency in the target language and bilinguals (i.e., most immigrants) against the test scores of native speakers, a distinction is usually made between verbal and nonverbal tests. Subtests with a substantial verbal component measure to an undesirable extent proficiency in the language of the test taken and underestimate the level of *g* of the tested nonnative speakers (see te Nijenhuis & van der Flier [1999] for a review of Dutch studies). The more limited the language skills, the larger the underestimate. Language bias plays a clear role in the testing of immigrants in Europe, but also in the testing of Blacks in South Africa, where the English used in the test, it is sometimes the second or even third language of the Black test taker.

In a study of Dutch immigrants, using a mixture of culture-loaded and culture-reduced tests te Nijenhuis and van der Flier (2003) found that the highly verbal subtest Vocabulary of the GATB is so strongly biased that it depresses the score on Vocabulary by 0.92 *SD*, leading to an underestimate of *g* based on GATB IQ, with as much as 1.8 IQ points due to this single biased subtest alone, whereas the other 7 subtests combined show only very little bias. However, one should not forget that subtests with a strong verbal component usually constitute only a small part of a test battery; due to the use of sum scores the strong bias in tests with a verbal component becomes diluted.

Looking at the effect of length of residence in the Netherlands on the scores on various intelligence tests also shows the influence of language. Tests without a verbal component show small to negligible correlations with length of residence, tests with a verbal component show moderate correlations, while language proficiency tests show large correlations (see te Nijenhuis & van der Flier, 1999; see van den Berg, 2001, p. 37). All these findings regarding the clear but modest role of language bias are in line with the findings of language bias when testing Hispanics who do not have a desirable level of proficiency in the target language or who are bilingual (Lopez, 1997; Pennock-Román, 1992).

In the US studies on Spearman's hypothesis it is usually native-born Blacks and Whites who are compared. Therefore language bias is not the problem that it is in the study of immigrants, and

Blacks and Whites in South Africa. However, studies of Hispanic immigrants may show language bias. In order to combine the diverse studies for a meta-analysis the effects of language bias had to be taken into account. We did this by leaving out subtests with a substantial language component for immigrants in Europe; Blacks in South Africa; and Mexican immigrants in the US for some studies. When there were still at least seven subtests left we recomputed the correlation between *d* and *g* and included that data point in the meta-analyses. Therefore Table 5 of Study 3 in some cases shows two correlations between *d* and *g*: one for all subtests and another after excluding one or more subtests with a substantial language component.

Research Questions. The first research question is: Which of the different subtests of the IQ batteries are language-biased? The second research question is: How large is the underestimation of IQ due to these language-biased subtests?

Method

The test of Spearman's hypothesis results in a scatter plot with a line which represents the regression of *d* on *g*. Language-biased subtests show larger values of *d* than is expected based on their *g* loading, so their data points are above the regression line. The distance of the data point from the regression line is a good measure of the language bias in the subtest. Language-biased subtests were identified and were left out of the analysis.

Results

What follows is a detailed description by IQ-test battery of the subtests we excluded from the analyses due to language bias or potential language bias. First, were the Dutch studies of majority group members and immigrants. In the Dutch RAKIT (Helms-Lorenz et al., 2003; te Nijenhuis et al., 2004; Tolboom, 2000) four subtests were identified as having a substantial language component: Verbal meaning, Learning names, Idea production, and Storytelling. Therefore these subtests were left out of the analysis. Figures 3-10 show several scatter plots of subtest patterns before and after language biased subtests were removed for several immigrant groups on the RAKIT. The Dutch IQ-test battery GATB (te Nijenhuis & van der Flier, 1997, 2005) includes a single subtest with a substantial language bias, namely Vocabulary, and therefore this subtest was left out of the analyses. Furthermore, in the Dutch DAT (te Nijenhuis et al., 2000) two subtests were identified as having a substantial language bias: Vocabulary and Language usage. These subtests were left out of the analyses.

Lynn and Owen (1994) used the JAT (Junior Aptitude Tests) to compare Indians and Blacks with Whites in South Africa. Three subtests were identified as having a substantial language bias: Reasoning, Synonyms, and Memory (paragraphs), and therefore were left out of the analyses.

Valencia and Rankin (1986) compared Mexican-Americans and Anglo-Americans on the K-ABC. The K-ABC (Kaufman Assessment Battery for Children) consists of ten mental processing subtests divided into a sequential (three subtests) and a simultaneous (seven subtests) processing scale. An achievement scale is also present on the K-ABC and consists of seven subtests that cover vocabulary, language development, general factual knowledge, mental arithmetic, and reading. Many of the K-ABC achievement scale subtests are commonly viewed on other tests as measures of verbal intelligence (Reynolds, 1994). Therefore these subtests can negatively influence performance for groups that have taken the test in a nonnative language. Since the Faces and Places, Arithmetic, Riddles, Reading Decoding, and Reading Comprehension subtests may show language bias, we omitted these five subtests from the analyses.

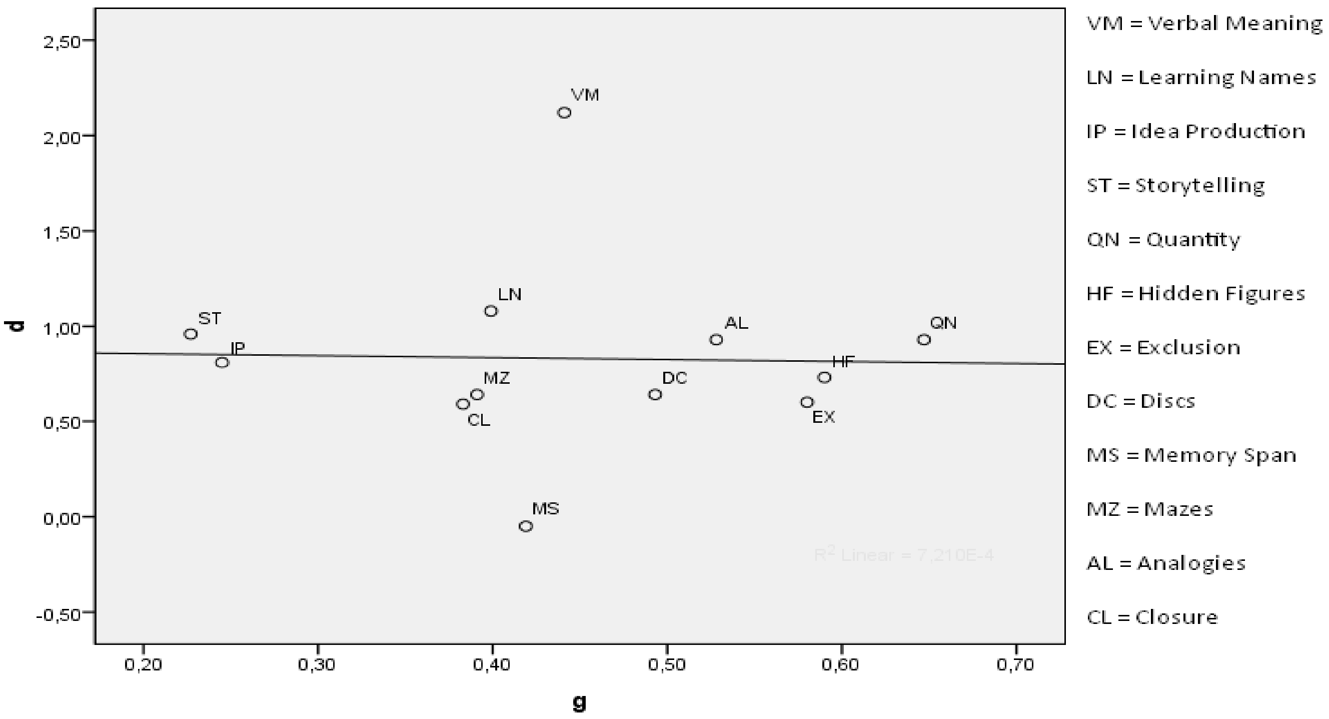

Figure 3

Scatter Plot of Subtest by g: Differences Between Dutch and Turkish Children of 7.8 Years on the RAKIT Including Language Biased Subtests; Data From Study by te Nijenhuis et al. (2004)

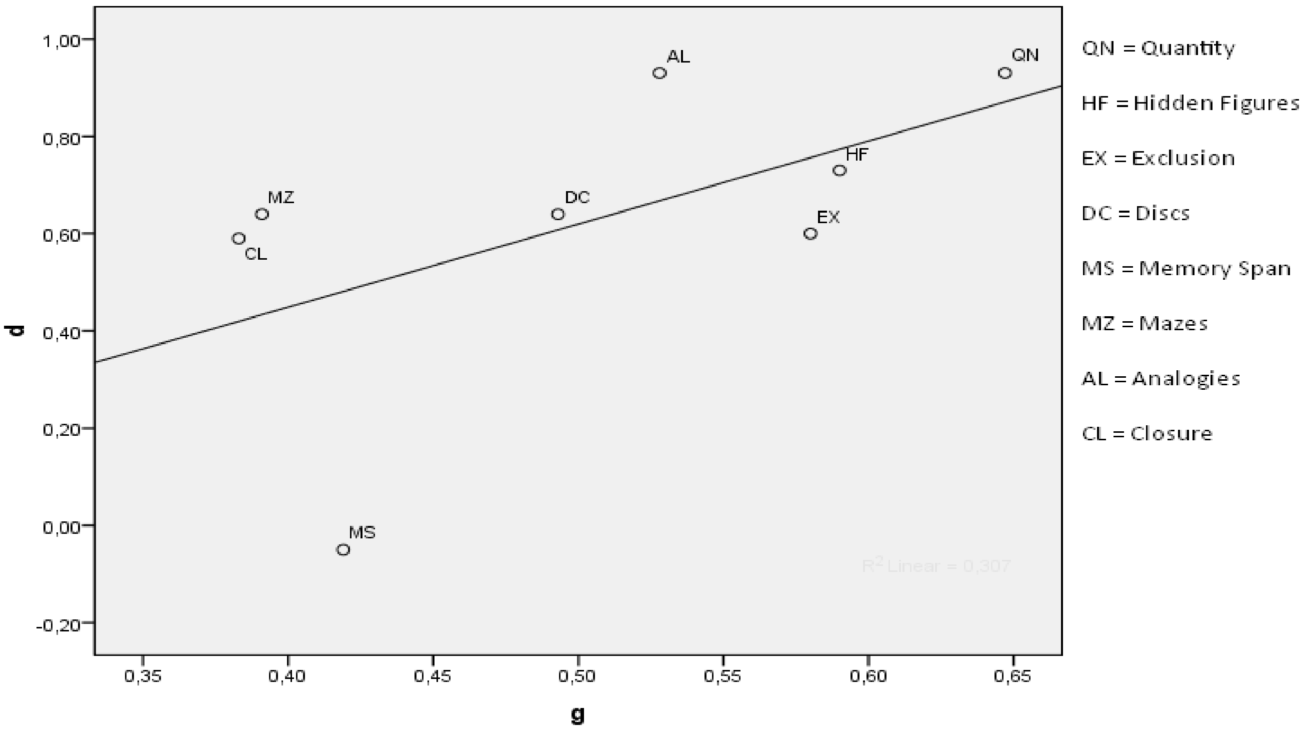

Figure 4

Scatter Plot of Subtest by g: *Differences Between Dutch and Turkish Children of 7.8 Years on the RAKIT Excluding Language Biased Subtests; Data From Study by te Nijenhuis et al. (2004)*

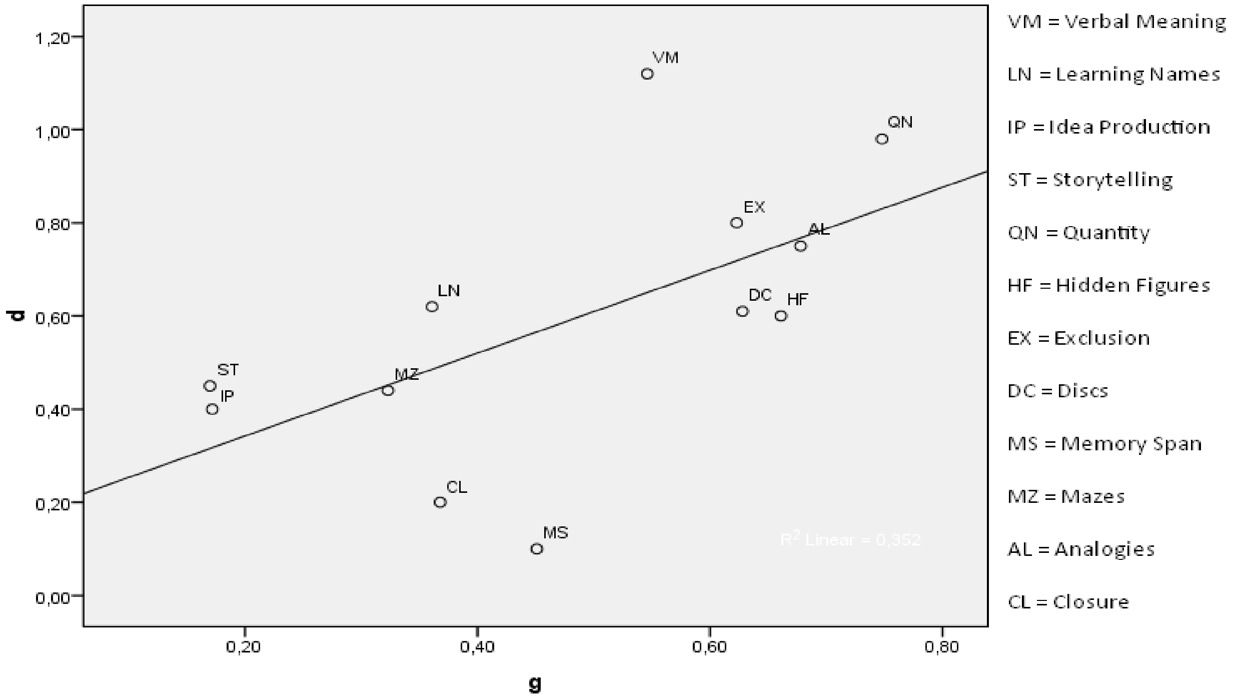

Figure 5

Scatter Plot of Subtest by g: Differences Between Dutch and Surinamese/Neth. Antillean Children of 9.8 Years on the RAKIT Including Language Biased Subtests; Data From Study by Tolboom (2000)

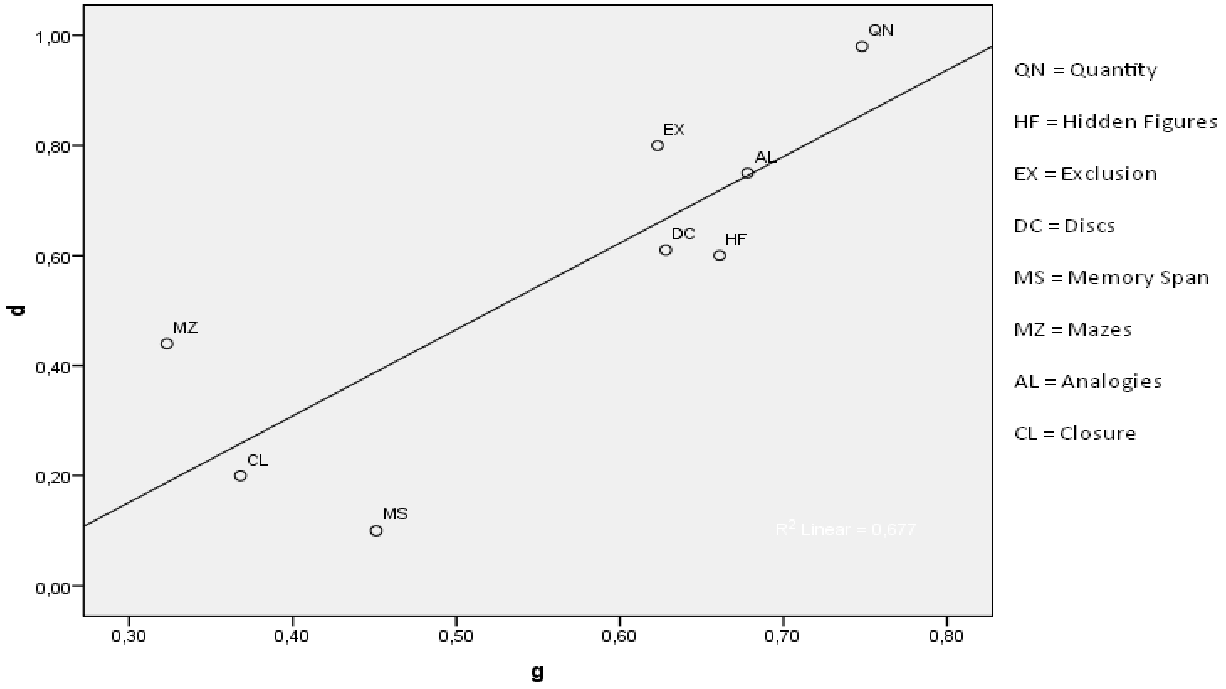

Figure 6

Scatter Plot of Subtest by g: Differences Between Dutch and Surinamese/Neth. Antillean Children of 9.8 Years on the RAKIT Excluding Language Biased Subtests; Data From Study by Tolboom (2000)

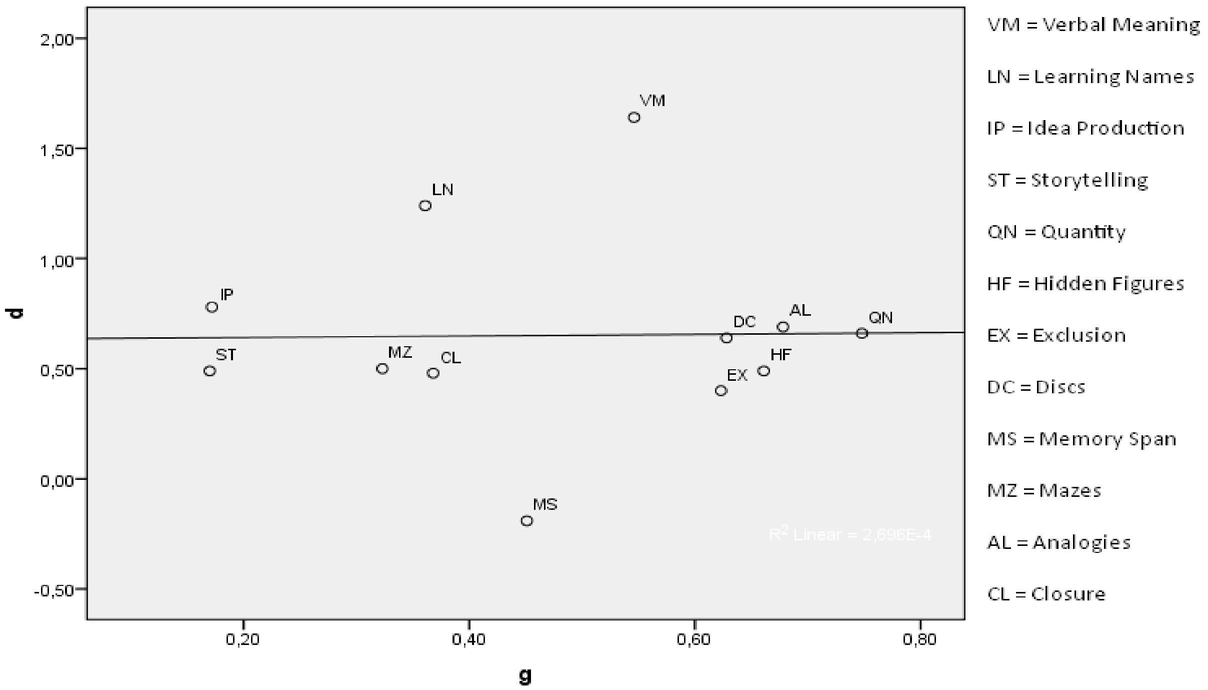

Figure 7

Scatter Plot of Subtest by g: Differences Between Dutch and Turkish Children of 9.8 Years on the RAKIT Including Language Biased Subtests; Data From Study by Tolboom (2000)

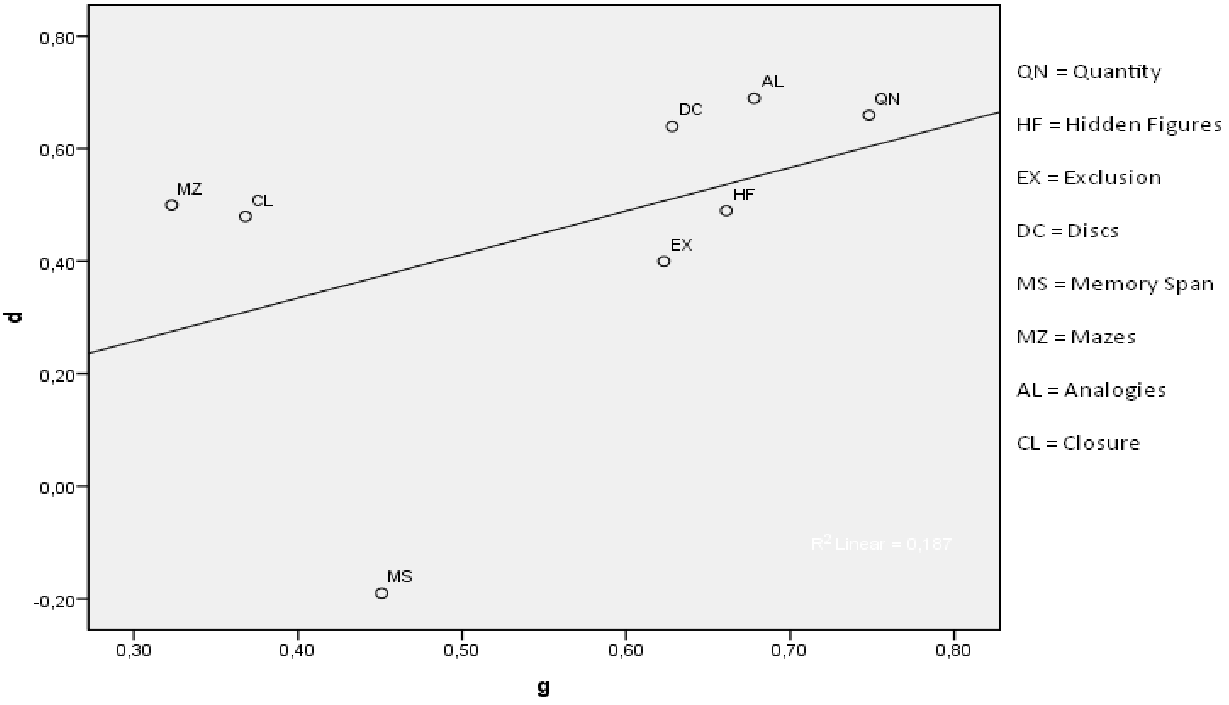

Figure 8

Scatter Plot of Subtest by g: Differences Between Dutch and Turkish Children of 9.8 Years on the RAKIT Excluding Language Biased Subtests; Data From Study by Tolboom (2000)

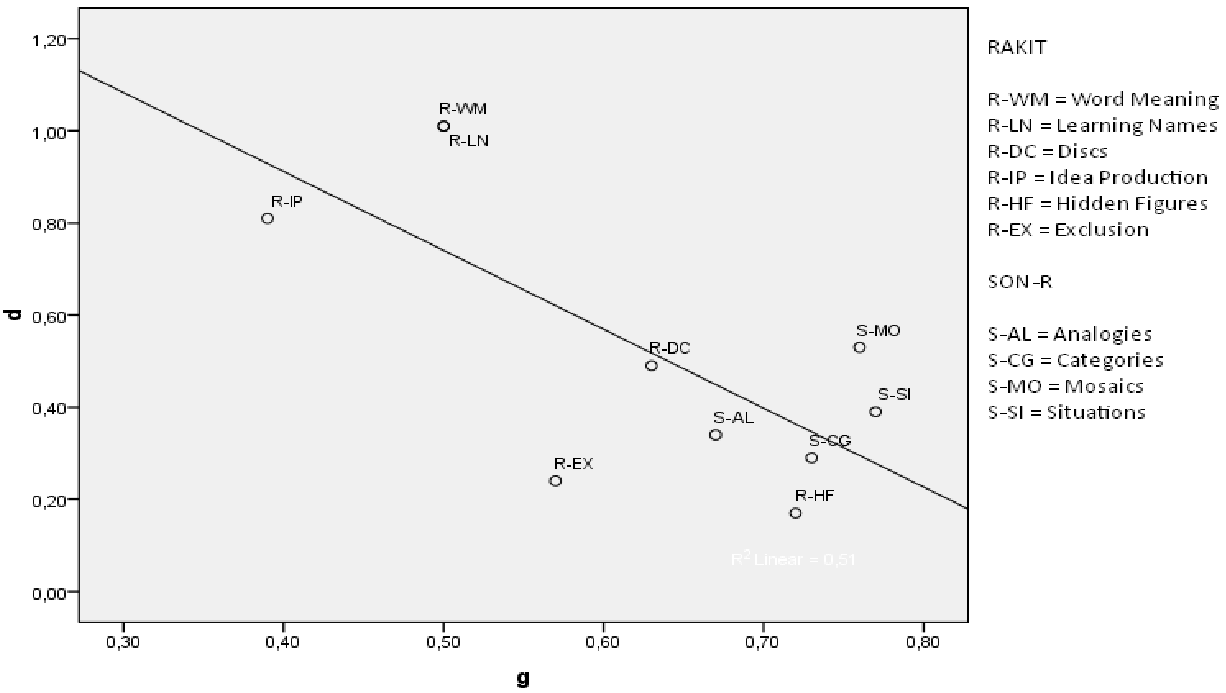

Figure 9

Scatter Plot of Subtest by g: Differences Between Dutch and Immigrant Children on the RAKIT and SON-R Including Language Biased Subtests; Data From Study by Helms-Lorenz et al. (2003)

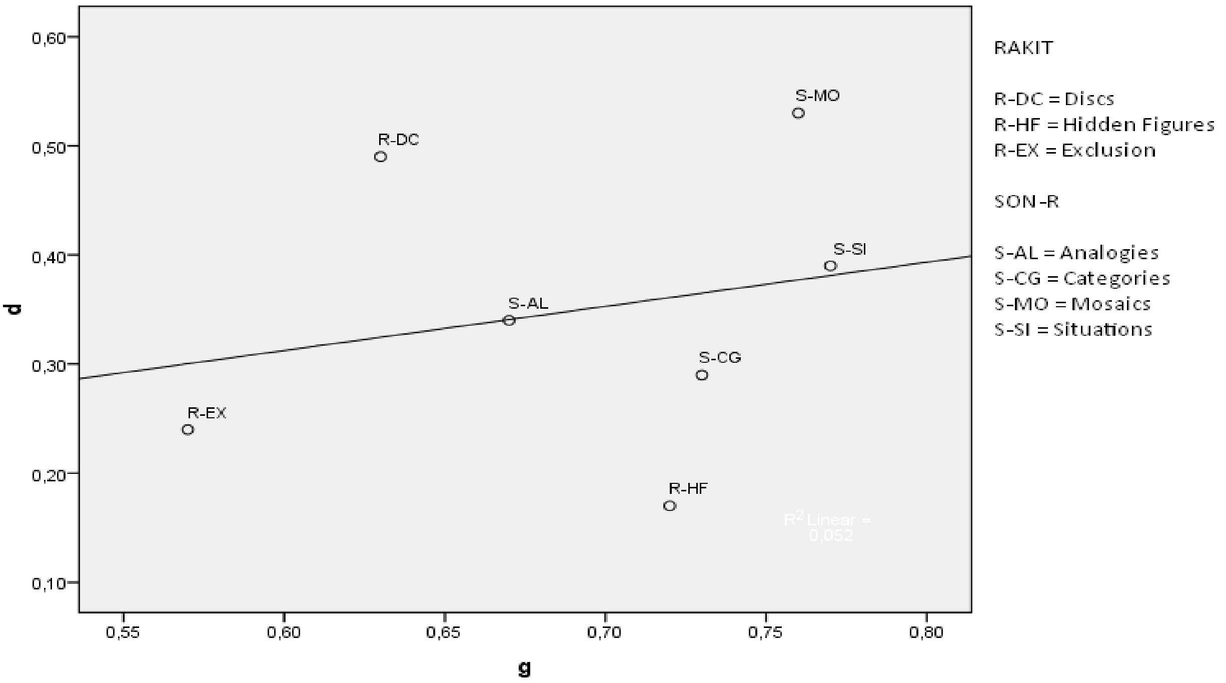

Figure 10

Scatter Plot of Subtest by g: Differences Between Dutch and Immigrant Children on the RAKIT and SON-R Excluding Language Biased Subtests; Data From Study by Helms-Lorenz et al. (2003)

Underestimation of IQ due to language bias

After having identified the subtests with a substantial language bias, the next step was to compute the degree to which these subtests disadvantage the people who do not have a desirable level of proficiency in the target language and bilinguals (i.e., most immigrants). Wicherts (2007) argued that a commonly used Dutch IQ test, the RAKIT, underestimates IQ of ethnic minority children about 7 points. Several IQ batteries contain subtests with language bias as we have shown above. The calculation of the underestimation of IQ due to language bias required several steps. First, after removing the language-biased subtests the new linear regression formula of *d* scores on *g* loadings was computed. This results in a regression line and a regression formula not distorted by language bias. Second, *g* loadings of the language-biased subtests of the IQ battery were separately entered in the new regression formula, resulting in a value of *d* for the data point expected solely on its *g*-loadedness, that is, without the influence of language bias. All these computed values of *d* were on the regression line. Third, this value was then subtracted from the *d* value still containing the language bias resulting in a value for the effect of language bias for this specific subtest. Fourth, the sum of these outcomes for all language-biased subtests in a specific battery was taken and then this sum was divided by the total number of subtests administered in the study, thereby also including the language-biased subtests. The result is an estimate, expressed in *SD*s, of how much the language-biased subtests depress the total IQ score of the battery in question. Table 4 shows the underestimation in IQ points for all the different test batteries in our study identified as comprising language-biased subtests.

Table 4

Underestimation of IQ in IQ Points for Groups on the Different IQ Batteries

study	IQ battery	number of language-biased subtests	group(s) tested	underestimation IQ (IQ points)
Te Nijenhuis, Tolboom, Resing, & Bleichrodt (2004)	RAKIT	4	S&NA	3.18
Te Nijenhuis, Tolboom, Resing, & Bleichrodt (2004)	RAKIT	4	Turks	4.56
Te Nijenhuis, Tolboom, Resing, & Bleichrodt (2004)	RAKIT	4	Mor.	3.00
Tolboom (2000)	RAKIT	4	S&NA	1.14
Tolboom (2000)	RAKIT	4	Turks	4.20
Tolboom (2000)	RAKIT	4	Mor.	2.75
Tolboom (2000)	RAKIT	4	S&NA	2.38
Tolboom (2000)	RAKIT	4	Turks	3.84
Tolboom (2000)	RAKIT	4	Mor.	2.68
Helms-Lorenz, van de Vijver, & Poortinga (2003)	RAKIT	3	Immigrants	2.74
Valencia & Rakin (1986)	K-ABC	5	Hispanics	10.03
Te Nijenhuis & van de Flier (1997)	GATB	1	Antilleans	.88
Te Nijenhuis & van de Flier (1997)	GATB	1	North. Afr.	1.19
Te Nijenhuis & van de Flier (1997)	GATB	1	Turks	1.8
Te Nijenhuis & van de Flier (2005)	GATB	1	Immigrants	1.14
Te Nijenhuis, Evers, & Mur (2000)	DAT	2	Immigrants	1.45
Lynn & Owen (1994)	JAT	3	Indians	-.648[1]
Lynn & Owen (1994)	JAT	3	Blacks	2.49

Note. S&NA = Surinamese and Netherlands-Antilleans; Mor. = Moroccans; North. Afr. = North Africans.

[1] The study by Lynn and Owen (1994) where Indians were tested shows a negative value of the underestimation which means that for this group there was no disadvantage because of language bias.

Conclusion and Discussion

When comparing different groups, language bias has to be taken into account, because the IQ of people who do not have a desirable level of proficiency in the target language is underestimated. However, this underestimation of IQ appears to be much smaller than the 7 points claimed by Wicherts (2007): the mean of the underestimation of IQ in the Dutch RAKIT samples is only 3.08 IQ points. The mean underestimation of IQ in for all studies in Table 4 is even lower, namely 2.71 IQ points. However, a clear exception is the Kaufman-ABC’s underestimate of the IQ of Hispanics with more than ten IQ points, a strong effect.

It is clear that when testing Spearman’s hypothesis language-biased subtests within a battery obscure the outcomes. For instance, the study by Helms-Lorenz et al. (2004) shows no support for Spearman's hypothesis. However removing the subtests with language bias may alter the author's conclusions. So, in the meta-analysis on Spearman’s hypothesis subtests with language bias were taken out.

Study 2: Correction for Deviation from Perfect Construct Validity

The deviation from perfect construct validity in *g* attenuates the values of *r* (*g* × *d*). In making up any collection of cognitive tests, we do not have a perfectly representative sample of the entire universe of all possible cognitive tests. Therefore any one limited sample of tests will not yield exactly the same *g* as another such sample. The sample values of *g* are affected by psychometric sampling error, but the fact that *g* is very substantially correlated across different test batteries implies that the differing obtained values of *g* can all be interpreted as estimates of a "true" *g*. The values of *r* (*g* × *d*) are attenuated by psychometric sampling error in each of the batteries from which a *g* factor has been extracted. We carried out a separate study to empirically estimate the values for this correction.

The more tests and the higher their *g* loadings, the higher the *g* saturation of the composite score is. The Wechsler tests have a large number of subtests with quite high *g* loadings, yielding a highly *g*-saturated composite score. Jensen (1998, p. 90–91) states that the *g* score of the Wechsler tests correlates more than .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 (1998. ch. 10) states that the average *g* loading of an IQ score as measured by various standard IQ tests lies in the +.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 tests' *g* score correlates about .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 .85 with "true" *g* loadings. As the Schmidt and Le (2004) 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* × *d*) after correction for the first four artifacts. To limit the risk of overcorrection, in previous studies a conservative choice of the value of .90 for the correction was made (te Nijenhuis, van Vianen, & van der Flier, 2007; te Nijenhuis, & Grimen, 2007; te Nijenhuis, de Pater, van Bloois, & Geutjes, 2009; te Nijenhuis & Franssen, 2009; te Nijenhuis & van der Flier, submitted). The observed correlation corrected 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 by using samples that took a large number of cognitive tests. Te Nijenhuis and Franssen (2010) analyzed Wechsler test data using a formula given by Jensen (1998, pp. 103-104) to compute the *g*-loadedness of a sum score

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

where r^2_{sg} = each subtest's squared *g* loading. The formula shows 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 results in a *g* loading of .92-95 for the various Wechsler Full Scale scores based on 10-12 subtests. It may be that having about 15 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, 6 subtests from a large collection of cognitive tests and 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 6 subtests with "true" *g*. Various combinations of 6 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 fifth correction when using a battery consisting of 6 cognitive tests.

Research Question. How strongly is the correlation between "true" *g* and the *g* from an artificial test battery a function of the number of subtests?

Method

We used two datasets with a large number of cognitive tests to create distributions 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 datasets.

Computations of g loadings. g loadings were computed using the same techniques as in the rest of this project. In general, using the full dataset, *g* loadings were computed by submitting a correlation matrix to a principal-axis factor analysis and using the loadings of the subtests on the first unrotated factor. In some cases *g* loadings were taken from studies where other procedures were followed; these procedures have been shown empirically to lead to highly comparable results.

Computation of g scores and "true" g scores. The various *g* score 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. The "true" *g* scores were the *g* scores computed on the full set of subtests. The other *g* scores were based on a smaller number of subtests.

Selection of tests for the analyses. The subtests used to create artificial test batteries had to be representative of subtests used in the batteries in most studies. This meant excluding subtests with *g* loadings below .30, such as certain simple reaction time tests. Combinations of test scores were also excluded when their *g* loadings were exceptionally high, by which we mean a *g* loading above .90, such as, for instance, the Wechsler Full Scale IQ score.

Combinations of subtests. It is rare to see combinations of three subtests described as a test battery, so we defined 4 subtests as the minimum required to create an artificial battery. The maximum number of subtests for an artificial battery was set at the total amount of subtests minus one in a specific dataset. *g* scores of these combinations were computed for all research participants using the *g* loadings computed on the full sample.

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*.

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.

Centers for Disease Control Data Set

Research participants

The U.S. Centers for Disease Control (CDC, 1988) provided an archival data set on 4462 males who had served in the United States Armed Forces (see Nyborg & Jensen [2000] for full details). The percentages of Whites and Blacks are 87% and 13%, which are almost exactly the percentages in the total United States population at that time. The sample consists of draftees and enlisted men and approximately half of the sample had served in Vietnam. The CDC's original purpose in obtaining these data was to assess the long-term effects of the veterans' military service some 17 to 18 years after induction. The total sample is fairly representative of the U.S. population in terms of race, education, income, and occupations. However, it should be noted that a mandate of the U.S. Congress excludes from military service all persons who score below the 10th percentile of national norms on a pre-induction general aptitude test. Therefore, the lower tail of the distribution of ability is somewhat truncated in this sample. There was no formal truncation at the top end of the scale. Self-selection and various other educational and social selective factors affecting enlistment or draft status might possibly result in some degree of underrepresentation of the potentially higher scoring individuals. The subjects' average age on entering the service in 1967 was 19.9 years (*SD* = 1.7); the average age at which they were tested by the CDC was 37.4 years (*SD* = 2.5).

Psychometric Variables

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

1. Grooved Pegboard Test (GPT) (Right Hand): A measure of manual dexterity and fine motor speed; the speed score is the reciprocal of the number of seconds taken to place a set of pegs in a grooved hole as quickly as possible.

2. GPT (Left Hand).

3. Paced Auditory Serial Addition Test (PASAT): A measure of mental control, mental speed, and computational and attentional abilities. The subject mentally adds a sequence of numbers in rapid succession; score is the total number of correct responses.

4. Rey-Osterrieth Complex Figure Drawing (CFD): Direct Copy score: A measure of visual-spatial ability and memory; the subject reproduces a complex spatial figure while the figure is in full view.

5. CFD, Copy from Immediate recall.

6. CFD, Copy from Delayed recall (20 minutes of other activities intervening).

7. Wechsler Adult Intelligence Scale-Revised (WAIS-R), General Information, scaled score.

8. WAIS-R, Block Design, scaled score.

9. Word List Generation Test (WLGT): A measure of verbal fluency; subject generates as many words as possible for 60 seconds that begin with each of three letters: F, A, S. Total number of words generated.

10. Wisconsin Card Sort Test (WCST): A measure of concept-formation, problem-solving, and set-switching abilities and use of feedback in decision making. Ratio of correct responses to countable responses.

11. Wide Range Achievement Test (WRAT): Measures ability to read aloud a list of single words (untimed). Total raw score.

12. California Verbal Learning Test (CVLT): A measure of verbal learning and memory; subject recalls a list of 16 words over five repeated learning trials. Total correct over 5 trials.

13. Army Classification Battery (ACB), Verbal Test administered at time of induction: A measure of verbal reasoning.

14. ACB, Verbal Test administered an average of 17 years after induction.

15. ACB, Arithmetic Reasoning Test, administered at time of induction.

16. ACB, Arithmetic Reasoning Test, administered an average of 17 years after induction.

17. Pattern Analysis Test (PAT): A visual spatial measure of pattern recognition, administered at induction.

18. General Information Test (GIT): Administered at time of induction.

19. Armed Forces Qualification Test (AFQT): A general aptitude battery; total score on four subtests (Word Knowledge, Paragraph Comprehension, Arithmetic Reasoning, Mathematics Knowledge). Administered at time of induction.

g loadings

The data from Table 1 are taken from Nyborg and Jensen (2000) and show the first of the three principal components extracted from the correlations among the 19 variables in the total sample in their study (*N* = 4,462). PC1 accounted for 44.6% of the total variance in the 19 tests and is interpreted as the general intelligence factor (*g*). The *g* loadings ranged from +.363 to +.856, with a median loading of +.697.

Table 1
Dataset From Centers for Disease Control: Tests and Their g Loadings

test	*g* loadings
GPT Right Hand	.363
GPT Left Hand	.371
PASAT	.599
CFD Copy	.510
CFD Immediate	.582
CFD Delayed	.581
WAIS-R Information	.775
WAIS-R Block Design	.697
WLGT	.531
WCST	.486
WRAT	.749
CVLT	.519
ACB Verbal *	.822
ACB verbal	.826
ACB Arithmetic *	.817
ACB Arithmetic	.824
PAT	.726
GIT	.710
AFQT	.856

Note. * Test administered at induction into armed forces. All other tests administered an average of 17 years after induction.

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 19 subtests and the AFQT sum score. The other *g* scores were based on a minimum of 4 subtests and a maximum of 18 subtests.

Selection of tests for the analyses

The Armed Forces Qualification Test (AFQT) yields a total score on four subtests (Word Knowledge, Paragraph Comprehension, Arithmetic Reasoning, Mathematics Knowledge) and has a g loading of .856, slightly higher than the *g* loading of ACB Verbal. So, the AFQT total score is included. The ACB Verbal and the ACB Arithmetic were taken twice, but with 17 years in between, so we treated them as separate test scores, and included them both.

Combinations of subtests

We choose at least 4 subtests as the criterion to create an artificial battery. The maximum number of subtests for an artificial battery was set at 18. Artificial test batteries were created by first taking subtests number 1, 2, 3, and 4; then taking subtests number 2, 3, 4, and 5; then subtests number 3, 4, 5, and 6, and so forth. *g* scores of these combinations were computed using the *g* loadings computed on the full sample.

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*.

Scatter plot

All 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. 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.

RAKIT and Dutch WISC-R Data Set

Research participants

Bleichrodt, Resing, Drenth, and Zaal (1987) carried out a study using both the Dutch WISC-R and the RAKIT. The RAKIT was given to a nationally representative sample of 1,415 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.

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 (Bleichrodt, Drenth, Zaal, & Resing, 1987):

1. *Closure*; the child is given very incomplete pictures and has to figure out the complete picture. According to Carroll's taxonomy Carroll (1993) 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 he (or she) 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 he 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 didn’t 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 respectively measures of Broad Retrieval Ability, Fluid Intelligence, and Crystallized Intelligence at stratum II.

The subtests of the Dutch WISC-R (van Haasen, de Bruyn, Pijl, Poortinga, lutje Spelberg, et al., 1986):

1. *Information*; the child has to answer verbally 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 is 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, such as: 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*; the child has to replicate using blocks 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 measures Broad Visual Perception at stratum II.

9. *Comprehension*; the child has to answer different questions in which they have to give their in sight and judgment about everyday life issues. This subtests 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 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 exactly 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.

g loadings

The data in Table 2 are taken from the RAKIT manual (Bleichrodt, Resing, Drenth, & Zaal, 1987, p. 142, Table 9.4).

Table 2

Combined Datasets of RAKIT and Dutch WISC-R: Tests and Their g Loadings

test		*g*
Dutch names	*English names*	
RAKIT		.44
Figuur Herkennen		.53
Exclusie	Exclusion	.47
Geheugenspan	Memory Span	.56
Woordbetekenis	Verbal Meaning	.39
Doolhoven	Mazes	.61
Analogieën	Analogies	.63
Kwantiteit	Quantity	.51
Schijven	Disks	.49
Namen Leren	Learning Names	.62
Verborgen figuren	Hidden Figures	.20
Ideeënproduktie	Idea Production	.25
Vertelplaat	Storytelling	.58
WISC-R		.45
Informatie	Information	.61
Onvolledige tekeningen	Picture Completion	.52
Overeenkomsten	Similarities	.57
Plaatjes ordenen	Picture Arrangement	.70
Rekenen	Arithmetic	.64
Blokpatronen	Block Design	.56
Woordenschat	Vocabulary	.43
Figuurleggen	Object Assembly	.27
Begrijpen	Comprehension	.35
Substitutie	Coding	.44
Cijferreeksen	Digit Span	.44
Doolhoven	Mazes	.53

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 5 subtests and a maximum of 23 subtests.

Selection of tests for the analyses

The subtests used to create artificial test batteries had to be representative of subtests used in the batteries in most studies. All subtests were used.

Combinations of subtests

Here we set 5 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 (1993) model that are most commonly represented in IQ batteries. They constitute subtests numbers 1, 2, 3, 4, and 11. 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 subtest12 and WISC-R subtest 1; WISC-R subtests 1 and 2; et cetera, 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.

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*.

Scatter plot

All 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. 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.

Computation of the correction value

The regression line of the scatter plot was used to estimate the correction values for the fifth correction of artifacts. First, the number of subtests after removing language-biased subtests was taken for every study. Second, the correction value per study was estimated by taking the cut-point of the regression line for every study based on the amount of subtests. Third, the weighted mean of these corrections per study was computed based on the harmonic mean of the groups used in the study. In this way one correction factor applicable for the correction for deviation from perfect construct validity for the entire sample was obtained.

Results Study 2: Correction for Deviation from Perfect Construct Validity

A scatter plot of correlations of artificial test batteries with “true” *g* against *N* should reveal that the larger *N* becomes, the higher the value of the correlation, with an asymptotic function between *r* and *N* expected. We checked to see which curve gave the best fit to the expected asymptotic function. Figure 1 shows the scatter plot of the correlation between two sum scores and sample size, and the logarithmic curve that fitted optimally for the data from the Center of Disease Control. Figure 2 shows the scatter plot of the correlation between two sum scores and sample size, and the logarithmic curve that fitted optimally for the data from the RAKIT and Dutch WISC-R.

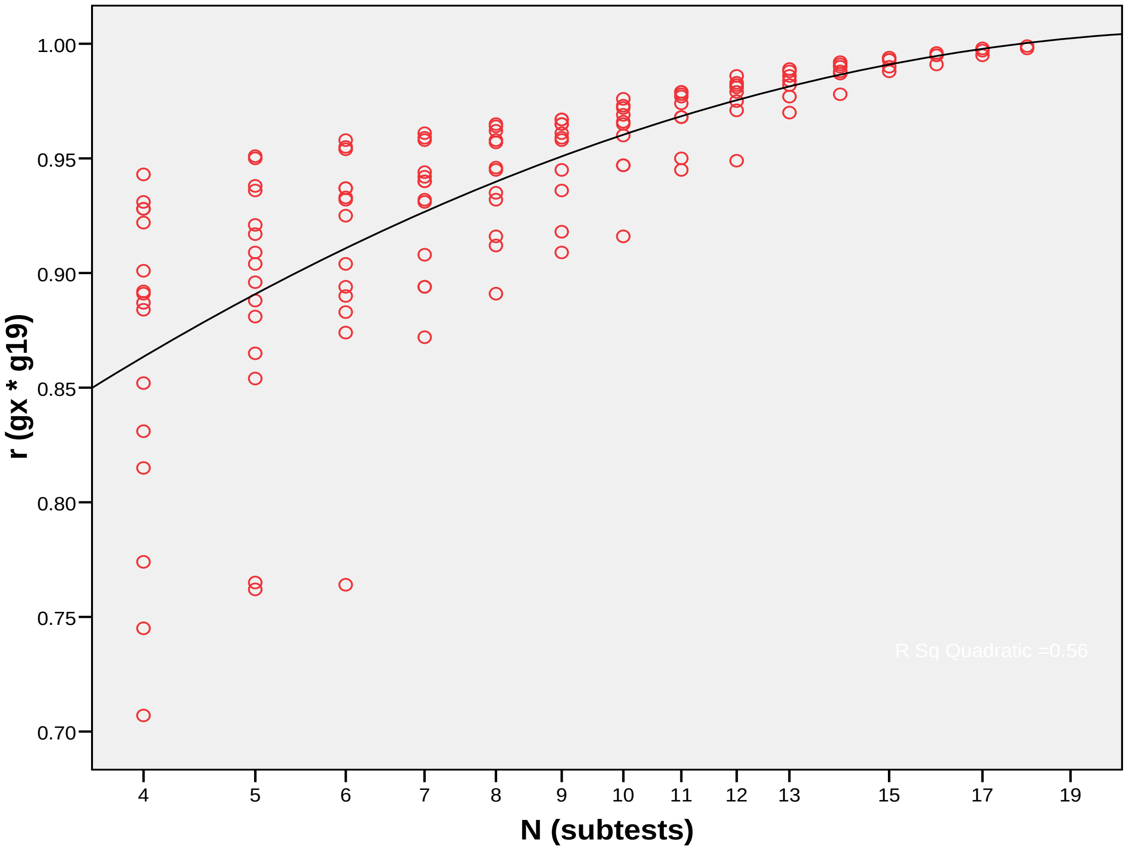

Figure 1

Centers for Disease Control: Scatter Plot of Correlation of g Score from Artificial Test Battery With “True” g and Sample Size; and Regression Line

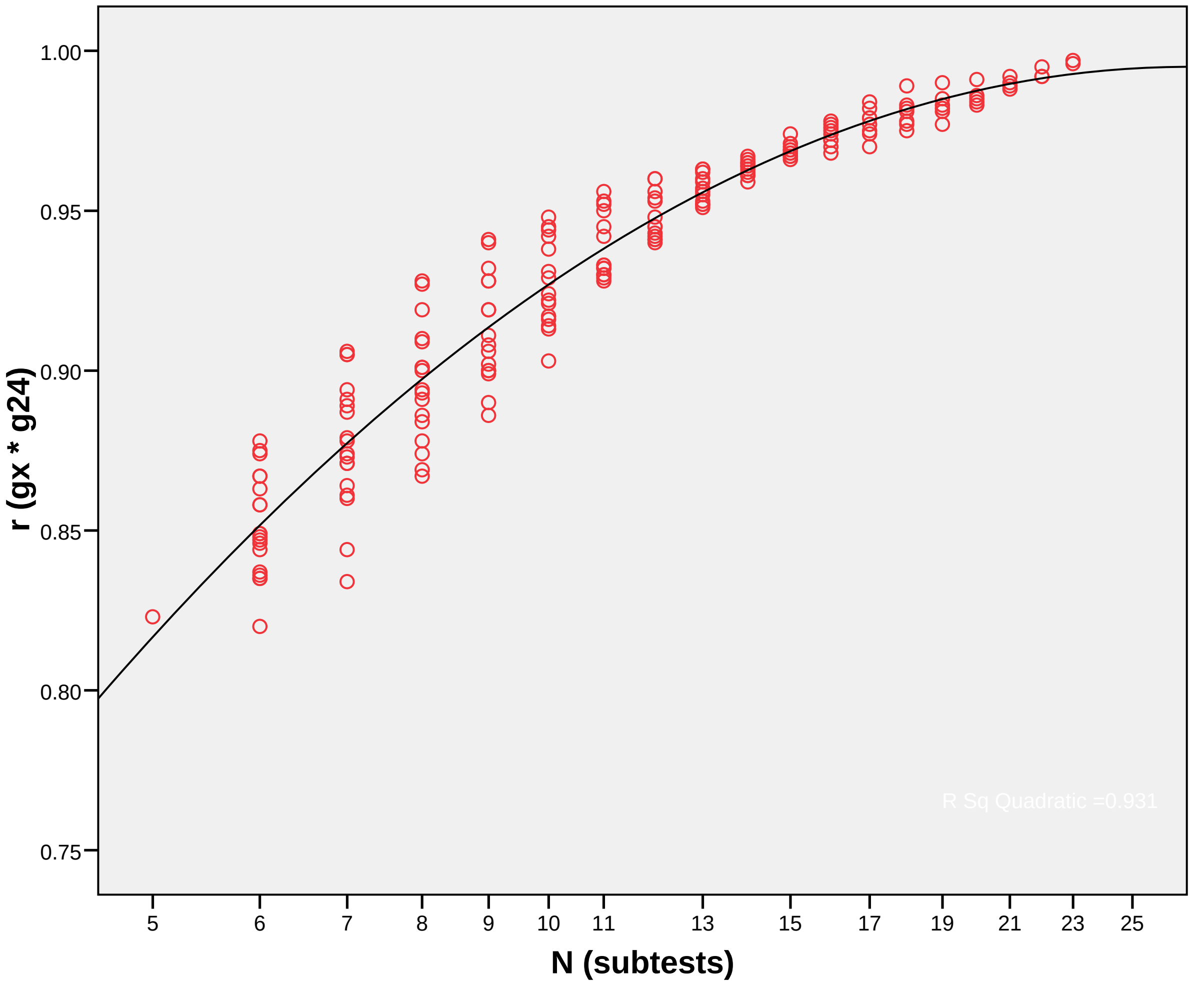

Figure 2

RAKIT and Dutch WISC-R: Scatter Plot of Correlation of g Score from Artificial Test Battery With “True” g and Sample Size; and Regression Line

Computation of the correction value

Table 3 shows the values of the correlations between *g* scores and ‘true *g*’ scores for every study based on the amount of subtests that were administered, not using the language-biased studies. For every study the correlations between *g* scores and ‘true *g*’ scores for the RAKIT and WISC-R dataset were estimated by taking the cut-point of the regression line based on the number of subtests. The weighted mean of the correlations across all studies was taken using the harmonic mean. This resulted in a value of .925 as a basis for the correction for imperfectly measuring the construct of *g*.

Table 3

Correction Values per Study Based on Number of Subtests

study	test battery	number of subtests[1]	N_{harmonic}	g x 'true g'
Jensen & Reynolds (1982) [2]	WISC-R	12	524	.95
Mercer (1984) [2]	WISC-R	12	643	.95
Nichols (1972) [2]	WISC-R	13	1,666	.95
Department of Defense (1982) [2]	ASVAB	10	3,247	.93
Kaufman & Kaufman (1983) [2]	K-ABC	13	608	.95
Kane & Brand (2008)	WISC-III	12	545	.95
Naglieri & Jensen (1987)	WISC-R, K-ABC	24	86	1.00
Jensen & Faulstich (1988)	WAIS-R	11	120	.94
Department of Labor [2]	GATB Aptitudes	8	3,120	.90
Hennessy & Merrifield (1976) [2]	CGP	10	697	.93
National Longitudinal Study [2]	CGP, SAT, ACT	12	3,348	.95
Nyborg & Jensen (2000)	Various tests	18	879	.99
Montie & Fagan (1988)	Stanford Binet 3 (LM)	12	86	.95
Hartmann, Kruuse, & Nyborg (2007)	Various tests	18	345	.99
Hartmann, Kruuse, & Nyborg (2007)	ASVAB	10	2,737	.93
Reynolds, Willson, & Ramsey (1999)	WISC-R	12	238	.95
Valencia & Rankin (1986)	K-ABC	8	100	.90
Te Nijenhuis, Tolboom, Resing, & Bleichrodt (2004) 7.8 years	RAKIT	8	77	.90
Tolboom (2000) 5.8 years	RAKIT	7	74	.88
Tolboom (2000) 9.8 years	RAKIT	8	74	.90
Helms-Lorenz, Van de Vijver, & Poortinga (2003)	RAKIT & SON-R	7	580	.88
Te Nijenhuis & Van der Flier (1997)	GATB	7	242	.88
Te Nijenhuis & Van der Flier (2005)	GATB	7	78	.88
Te Nijenhuis, Evers, & Mur (2000)	DAT	7	165	.88
Rushton (2001)	WISC-R	12	174	.95
Rushton & Jensen (2003)	WISC-R	10	203	.93
Lynn & Owen (1994)	JAT	7	1,070	.93

Note: [1] After removing subtests with language bias. [2] These studies were taken from Jensen (1985)

Conclusion and Discussion

Where previous studies (te Nijenhuis, van Vianen, & van der Flier, 2007; te Nijenhuis, & Grimen, 2007; te Nijenhuis, de Pater, van Bloois, & Geutjes, 2009; te Nijenhuis & Franssen, 2009; te Nijenhuis & van der Flier, submitted) used a conservative value of .90 as a basis to limit the risk of overcorrection, this new method for computing the correction for imperfectly measuring *g* shows that the correction used before was too strong. These researchers applied a correction of ten percent to compute rho-5, but in this paper we estimated the value of the correlation between g scores and "true g" scores to be .925. We rounded of the value for the correction of the fifth artifact to 7.5 % for the computation of rho-5. Due to time restrictions this correction was applied to all datasets; a specific correction value for each study was not computed.

Study 3: Psychometric meta-analysis of Spearman's hypothesis

Spearman’s hypothesis states that the different relative magnitudes of the Black/White differences on various tests are a function of each test’s *g* loading. This hypothesis has since been tested in numerous studies in the US, Europe, Asia, and Africa. However, a meta-analysis on this topic had not previously been conducted. Therefore, in this paper we report the results of a psychometric meta-analysis of Spearman's hypothesis.

Research Questions

The first research question is: In general how strong is the true correlation between group differences and *g* loadings? The second research question is whether the correlation between group differences and *g* loadings differs by group.

Method

First the classical method for testing Spearman’s hypothesis, as reported by Jensen, is described. Then we describe the psychometric meta-analytical approach.

The classical method for testing Spearman's hypothesis

Jensen (1993) 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 >.85.

4. The variables should measure the same *g* in the different groups; the congruence coefficient of the *g* values should be >.95.

5. The *g* loadings of the variables should be determined separately in each group. If the congruence 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.

Because we carried out a psychometric meta-analysis, only the first five requirements applied. Requirement 6 was replaced by meta-analytical corrections for reliability, so we started with the correlation before the correction for attenuation was applied and then applied the meta-analytical correction. Requirement 7 does not apply in a PMA: Total sample sizes becomes so large that significance testing is a waste of time.

General inclusion and decision rules

For a study to be included in the meta-analyses and in the exploratory studies, three criteria had to be met: First, in order to obtain a reliable estimate of the true correlation between the two variables the cognitive batteries had to have a minimum of seven subtests. Second, the test had to be well-validated. Third, since studies with a test-retest effect would influence the 'true' correlation between *d* and *g* (see discussion below) – they were excluded. That is, studies using a counterbalanced design and the scores of the re-administration of an IQ battery within a test-retest design were omitted from the analysis. (In a counterbalanced design, participants are administered two IQ batteries, X and Y, in different orders. Half of the participants take test X first, then test Y, and vice versa).

Specific criteria for inclusion

Seven specific decision rules for selecting relevant studies for this meta-analysis were used. First, only studies reporting a test of Spearman's hypothesis, computing a correlation between standardized group differences and *g* loadings, were included in the analysis. For example, Peoples, Fagan, and Drotar's (1995) study compared African-Americans with European-Americans on intelligence scores, but they did not test if the difference was due to *g*. Therefore that study was not included. An exception was made for the Vernon and Jensen (1984) study in which the correlation between *g* loadings and Black/White differences is not explicitly reported. However, it is obvious that this study includes a test of Spearman's hypothesis, so it is included in the meta-analysis. A second exception was made for the study by Ja-Song and Lynn (1992) where a test of Spearman's hypothesis was not explicitly conducted. However, it is obvious that this study is part of a series on Spearman's hypothesis conducted by Lynn, Jensen, and others and it is therefore included in our Table of miscellaneous studies. A very large study in the armed forces (Carretta & Ree, 1995a; Carretta & Ree 1995b; Carretta 1997) reported both *d* and *g* but in different articles and no correlation was computed, and therefore it too was not included in the present meta-analysis.

Second, in a few cases the same data were reported in more than one study. When multiple studies used the same sample, the sample with the largest *N* was included in the meta-analysis and the others were excluded from further analysis. For example, the study from Reynolds and Gutkin (1981) used a subsample of the WISC-R national standardization, whereas the whole sample was used by Jensen and Reynolds (1982), so the latter study was included.

Third, some studies were included in the meta-analysis where a number of subtests from different intelligence batteries had been used. If the subtests met the first five requirements stated by Jensen (1993) and there were seven or more subtests, the studies were included in our analysis. The study by Naglieri and Jensen (1987) used two IQ batteries, both having more than seven subtests (WISC-R: 11 subtests, K-ABC: 13 subtests) that were administered to the same group. We decided to combine all the subtests to create one data point for our meta-analysis so as to get the most reliable correlation between standardized group difference scores (*d*) and *g* loadings.

Fourth, most of the Spearman's hypothesis studies used *d* and *g* scores based on a number of single subtest scores to compute a correlation. However, some studies reported scores based on a combination of several subtests. For example, the studies by Nyborg and Jensen (2000), and Hartmann et al. (2007) reported *d* and *g* scores based on a total test score (AFQT), and these were included by these authors to compute *r.* In our meta-analysis we included only reported scores based on a maximum of two subtests, so the correlations between *d* and *g* were recomputed excluding the AFQT score for these two studies. In this way all correlations are computed in the same way across the different studies. Moreover, studies reporting aptitude scores based on a sum score of two subtests can be included in the meta-analysis (i.e., Department of Labor data, reported by Jensen, 1985).

Fifth, members of a specific group had to be representative of their group. Hennessy and Merrifield (1976) partialed socioeconomic status (SES) out of their analysis because they were primarily concerned with ethnic population differences in the factor structure of abilities. Therefore we did not reason not include this study in the present meta-analysis.

Sixth, the study by Nagoshi, Johnson, DeFries, Wilson, and Vandenberg (1984) in Hawaii was not included in the analysis because the mean subtests scores of the different groups could not be obtained. After a thorough search on the internet and several databases for articles which reported this necessary information with no results we contacted Dr. Craig Nagoshi and Dr. John DeFries. They replied that due to an agreement signed by the co-investigators prior to data collection, which stated that ethnic group means could not be published this information was never reported in any article. Therefore we were forced to leave this large study out of our analysis.

Seventh, in the analyses on intelligence batteries we focused on subtests from classical intelligence batteries. The study by Helms-Lorenz et al. (2001) included two reaction time measures, so we left them out of our analysis.

Corrections for Artifacts

Psychometric meta-analytical techniques (Hunter & Schmidt, 1990, 2004) were applied using the software package developed by Schmidt and Le (2004). Psychometric meta-analysis is based on the principle that there are artifacts in every dataset and that most of these artifacts can be corrected. In the present meta-analyses we corrected for five artifacts identified by Hunter and Schmidt (1990) that alter the value of outcome measures. These are: (1) sampling error, (2) reliability of the vector of *g* loadings, (3) reliability of the vector of a specific variable of theoretical interest (4) restriction of range of *g* loadings, and (5) deviation from perfect construct validity. In the present exploratory studies, using bare-bones meta-analytical techniques, we corrected for only one artifact, namely sampling error.

Correction for Sampling error. 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.

Correction for Reliability of the Vector of g Loadings. The values of *r* (*d* x *g*) are attenuated by the reliability of the vector of *g* loadings 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. Several samples that differed little on background variables were compared. 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 in the age of 6 to 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 we compared samples in the age of 18 to 95 years.

te Nijenhuis and Grimen (2007) collected correlation matrices from test manuals, books, articles, and technical reports. The large majority came from North America, with a large number from European countries, and also a substantial number from Korea, China, Hong Kong, and Australia. This resulted in about 700 data points, which yielded 385 comparisons of *g* loadings of comparable groups from which to estimate the reliability for that group. To give an illustration of the procedure, van Haasen et al. (1986) report correlation matrices of the Dutch and the Flemish WISC-R for 22 samples in the age of 6-16 years. We compared samples of children in the age of 6 to 17 years with other samples of children who do not differ by more than 1.5 years. Because the samples of children reported in van Haasen et al. (1986) were between 6 and 17 years we only compared children who did not differ more than 1.5 years. The *N*s in these samples were comparable. The resulting average correlation was .78 (combined *N* = 3,018; average *N* = 137).

A scatter plot of reliabilities against *N*s should show that the larger *N* becomes, the higher the value of the reliability coefficients, with an asymptotic function between *r* (*g* x *g*) and *N* expected. We checked to see which curve gave the best fit to the expected asymptotic function. The logarithmic regression line resembled quite well the expected asymptotic distribution for reliabilities. However, because the extreme range on the x axis resulted in a picture that is not informative, the regression line for r (g x g) and N is not reported. For the same reason we divided Figure 11 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.

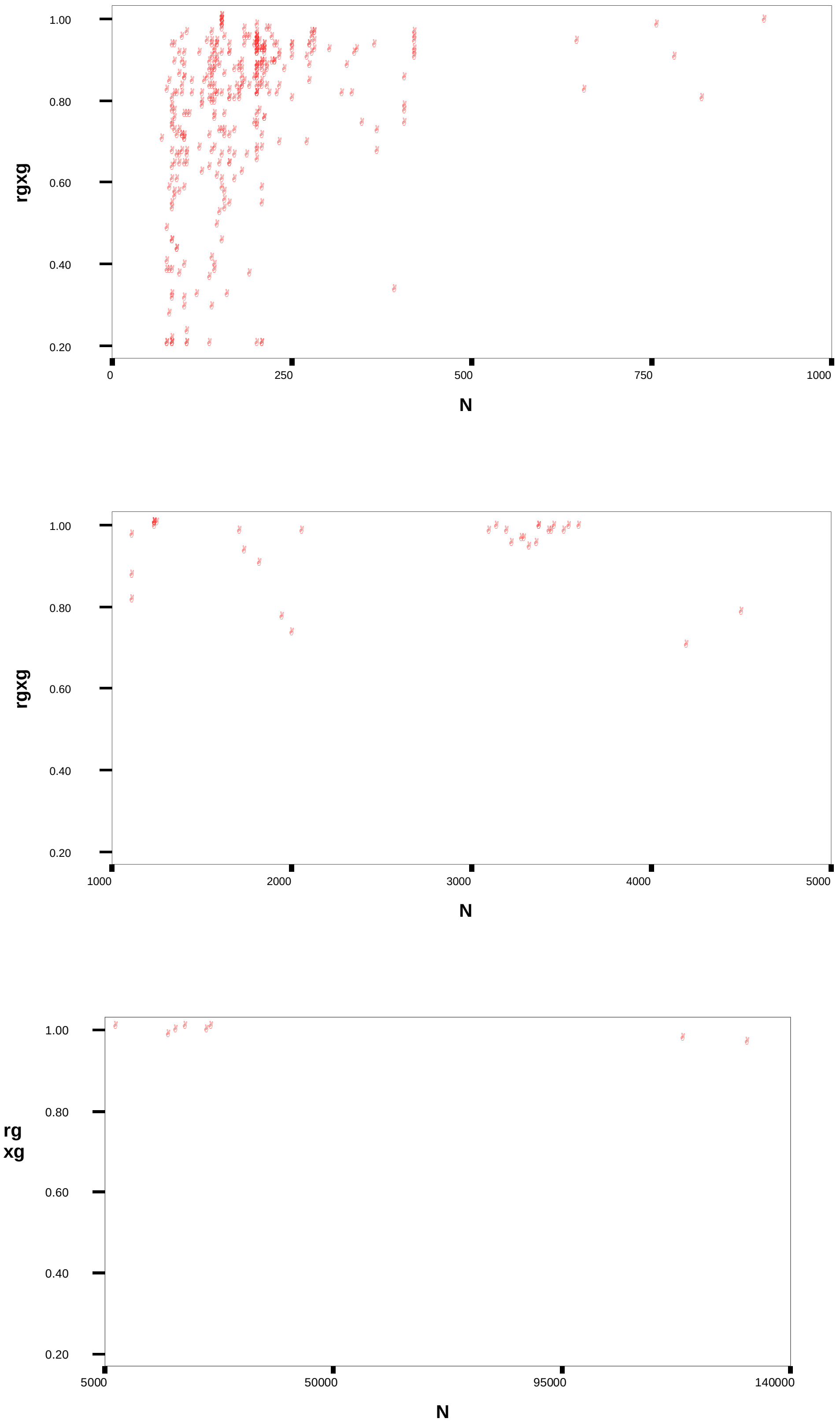

Figure 11

Three Scatter Plots of Reliability of the Vector of g Loadings and Sample Size Each for Different, Adjacent Ranges of N

Correction for Reliability of the Vector of the Second Variable. The values of *r* (*g* × *d*) are attenuated by the reliability of the *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 vector of group differences 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.

Correction for Restriction of Range of g Loadings. The values of *r* (*g* × *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 (1998, 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 (1990, pp. 47-49) state 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 $u = SD_{\text{study}}/SD_{\text{ref}}$.

First, as references we 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 Bellevue (W-B), Wechsler Preschool and Primary Scale of Intelligence (WPPSI), Wechsler Intelligence Scale for Children (WISC), Wechsler Intelligence Scale for Children–Revised (WISC-R), Wechsler Intelligence Scale for Children–Third Edition (WISC-III), and the Wechsler Intelligence Scale for Children–Fourth Edition (WISC-IV) from datasets from countries all over the world was 0.132. We used this value as our reference in the studies with children. The average standard deviation of g loadings of the various versions of the Wechsler Adult Intelligence Scale (WAIS), Wechsler Adult Intelligence Scale–Revised (WAIS-R), and the Wechsler Adult Intelligence Scale–Third Edition (WAIS-III) from datasets from countries all over the world was 0.107. This was used as the reference value in the studies with adults. 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.

Secondly, values of *u* using a different reference group were calculated to test the effect of the choice of the reference group on the value of rho-4 and the percentage variance explained. As the second reference we used the *SD* of the *g* loadings of the Dutch GATB based on a sample size of 500 (te Nijenhuis & van der Flier, 1997). However, the participants took 47 IQ tests and therefore a highly reliable estimate of *g* can be made. We estimated the equivalent to be *N* = 2500. The GATB has a much larger range in *g* loadings than the Wechsler scales. A couple of studies with large samples took the GATB (Department of Labor, reported in Jensen, 1985; te Nijenhuis & van der Flier, 1997; te Nijenhuis & van der Flier, 2005) so they have a strong influence on the rho corrected for restriction of range. Finally, the outcomes of the two analyses using different IQ batteries as references were compared. The Hunter and Schmidt meta-analytical program computes only the aforementioned four corrections. We will refer to the observed correlation corrected for sampling error, unreliability of the vector of *g* loadings and the second vector, and range restriction as rho-4.

Correction for Deviation from Perfect Construct Validity. Where previous studies (te Nijenhuis, van Vianen, & van der Flier, 2007; te Nijenhuis, & Grimen, 2007; te Nijenhuis, de Pater, van Bloois, & Geutjes, 2009; te Nijenhuis & Franssen, 2009; te Nijenhuis & van der Flier, submitted) used a conservative value of .90 as a basis to limit the risk of overcorrection, the new method for computing the correction for imperfectly measuring *g* from Study 2 shows that the correction used before was too strong. These researchers applied a correction of ten percent to compute rho-5, but in this paper we estimated the value of the correlation between *g* scores and "true *g*" scores to be .925. We rounded of the value for the correction of the fifth artifact to 7.5 % for the computation of rho-5. Due to time restrictions this correction was applied to all datasets; a specific correction value for each study was not computed.

Searching and screening studies

To identify studies for inclusion in the meta-analysis, both electronic and manual searches were conducted for studies that contained cognitive ability data of different racial groups and in which Spearman's hypothesis was tested. Three methods were used to obtain scores for different groups from published studies for the present meta-analysis. First, an electronic search for published research using PsycINFO, PiCarta, Academic search premier, Web of science, and PubMed was conducted. The following combinations were used to conduct the searches: any keyword that contains the words "Spearman's hypothesis", or "Jensen effect(s)". Suggestions have been offered to replace the term Spearman's hypothesis by the term Jensen effect (Rushton, 1998) but the new term has not caught on. Second, reference lists of all currently included empirical studies were checked to identify any potential articles that may have been missed by earlier search methods. Finally, several well-known researchers who have conducted research on Spearman's hypothesis were contacted in order to obtain any additional articles or supplementary information. This resulted in 44 studies with 79 correlations of which 57 correlations were used in the present study.

Estimating d and g

A general aim of the meta-analysis was to get the best estimates of *d* and *g* as possible. Some of the original studies used small sample sizes to compute *g*, whereas *g* loadings based on much larger sample sizes were available in the literature. So, in those cases we substituted the *g* loadings reported in the original study with the *g* loadings based on much larger sample sizes. With concern to quality of datasets we looked both at sample size and representativeness of the populations. When samples were representative but rather small and at the same time there was a very large non-representative sample we generally preferred the very large sample. Many American samples consisted of a large number of Whites, and a much smaller number of Blacks or Mexican-Americans. In most cases we choose the *g* loadings based on the large White samples. In some cases *g* loadings were only reported for the combined samples of Blacks and Whites and the resulting *g* loading was used. However, in the South African studies the situation was different. For instance, for the South African JAT correlation matrices are reported for four large groups, so we computed *g* loadings for each of the four groups and then computed the average *g* loading.

Computation of score differences between two different groups

Score differences between two groups (*d*) were computed by subtracting the mean of the lower scoring group from the mean of the higher scoring group (to generally obtain positive scores) and then dividing the result by the *SD* of the standardization group. The standardization group scores were obtained from the manual of the IQ battery and were computed by means of a weighted average score to match the age range of the participants as closely as possible. When *SD* scores of the standardization group could not be obtained, *SDs* of the largest group in the study, usually the majority group, were taken. Scores of the largest group are the most reliable ones. Jensen (1985) used a different technique: He expressed the average differences between groups in standardized units by taking the weighted average standard deviation within groups, the weights being the respective sizes of the two samples. All of the background information for a couple of Jensen's datasets (i.e., the data from the Department of Labor and from the Department of Defense) was not available to us, so values for *d* were taken directly from Jensen (1985). However, this has the disadvantage that the values of the *SD,* using the same IQ battery, vary by study.

g Loadings

g Loadings were obtained in the same way as the *d* scores when the manual of the IQ battery was available. *g* Loadings matching the age range of the group participants were used to compute *r* (*d* x *g*). Therefore, the weighted average *g* loadings were computed, matching the age range of the participants to the age range of the *g* loadings as close as possible. When the manual of the IQ battery was not available or if it did not contain information on *g* loadings, *g* loadings of the group with the biggest *N* was used to compute *r.* Two studies of *g* loadings had an *N* of 500, but the participants took 47 IQ tests. Therefore a highly reliable estimate of *g* can be made. We estimated the equivalent to be *N* = 2,500.

Correction for Sampling Error

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. Most of the groups compared were not of equal size and in some comparisons one group was much smaller than the other. Therefore for all comparisons we computed harmonic means for sample size using the following formula (Klockars & Sax, 1987), where *n* is the number of scores and *x* is an individual score:

$$\frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} + \ldots + \frac{1}{x_n}}$$

General buildup of the analyses

The data points from the studies were divided into five categories: IQ batteries, Raven's Matrices at the item level, reaction time measures, educational outcomes, and a category for the results of miscellaneous studies. In the first analysis of IQ batteries, the studies from the U.S., the Netherlands, and Africa were combined and the overall rho was computed. Subsequently, separate meta-analyses of IQ batteries were carried out for the US, the Netherlands, and for Africa, respectively. For educational and training outcomes there were seven studies (eleven correlations) which reported a test of Spearman's hypothesis, so for these studies a bare-bones meta-analysis was carried out. For the studies which used the Raven's Progressive Matrices we simply report the correlations between *d* and *g* in a Table. Due to time constraints the reaction-time studies were not analyzed.

Psychometric meta-analytical techniques (Hunter & Schmidt, 1990, 2004) were applied to the resulting thirty-eight *r* (*g* × *d*)'s using the software package developed by Schmidt and Le (2004). In the present study we corrected for the five artifacts (mentioned above) that alter the value of outcome measures listed by Hunter and Schmidt (1990).

Results Study 3: Psychometric meta-analysis on Spearman's hypothesis

The results of the studies on the correlation between *g* loadings and the score differences between groups (*d*) are shown in Table 5. The Table gives data derived from twenty-six studies, comprising a number of thirty-eight data points, with participants numbering a total of 67,715. The Table also lists the reference for the study, the cognitive ability test used, the groups that were compared, the correlation between *g* loadings and *d*, the harmonic mean, and the mean age (and range of age). It is clear that the large majority of the correlations are strongly positive.

Wechsler test batteries as a standard for restriction of range

Table 6 presents the results of the psychometric meta-analysis of the thirty-eight data points where the Wechsler test batteries have been used as the standard for the correction for restriction of range. It shows (from left to right): the number of correlation coefficients (*K*), total sample size (*N*), the mean observed correlations (*r*) and their standard deviation (SD_r), the correlations one can expect once artifactual error from unreliability in the *g* vector, the *d* vector, and range restriction in the *g* vector have been removed (rho-4), and their standard deviation ($SD_{\text{rho-4}}$), and the true correlation one can expect when corrections for all five artifacts have been carried out (rho-5). The next two columns present the percentage of variance explained by artifactual errors (%VE), and the 80% confidence interval (80% CI). This interval denotes the values one can expect for rho-4 in sixteen out of twenty cases.

The analysis of all 38 data points yields an estimated correlation (rho-4) of .57, with only 24% of the variance in the observed correlations explained by artifactual errors. However, Hunter and Schmidt (1990) state that extreme outliers should be left out of the analysis, because they are most likely the result of errors in the data. They also argue that extreme outliers artificially inflate the *SD* of effect sizes and thereby reduce the amount of variance that artifacts can explain.

There are statistical reasons and theoretical reasons to exclude outliers and extreme outliers. A first statistical reason for exclusion is when a data point falls several standard deviations below the mean of the sample of data points without the outliers. A second statistical ground to exclude a data point is when the distribution of data points in the scatter plot is highly uneven, meaning there are large gaps between adjacent data points. A theoretical reason to exclude a data point is when a dataset is dissimilar from all other datasets, for instance when the group in question is quite different from the groups in the other studies. The strongest case for excluding data points is when there are both good statistical and theoretical reasons.

Figure 12 shows the scatter plot of all correlations *r* (*d* x *g*) against the harmonic mean. We choose to first leave out three extreme outliers, with a value of *r* more than 10 *SD* beneath the average *r* of the final sample of thirty-five data points. The studies by Department of Defense (1982) and Lynn and Owen's (1994) data on Whites/Indians, and Whites/Blacks were considered extreme outliers and therefore omitted from the analysis. These are studies with large sample sizes, so meta-analytical theory predicts a high correlation and not the small correlations reported by the authors of the studies. Removing these data points resulted in a substantial change in the value of the correlation (rho-4), a large decrease in the *SD* of rho-4, and a huge increase in the amount of variance explained in rho-4 by artifacts: 77 % of the variance is now explained. We also checked what would happen when instead of three, six outliers were removed. The studies by Jensen and Faulstich (1988), Valencia and Rankin (1986), and Tolboom's (2000) data on 5.8-year old Dutch and Moroccan testess were excluded because there was a huge gap between these studies and the adjacent data points (see Figure 12). The study by Jensen and Faulstich (1988) investigated White and Black prisoners and they even found a negative correlation between *d* and *g.* The value of rho-4 did not change drastically after removing another three data points, but the *SD* of rho-4 decreased to a value of 0.0, and the percentage of variance explained by artifacts increased to 130. Finally, a correction for deviation from perfect construct validity in *g* took place, using the value of 7.5 %. This resulted in a value of .71 for the final estimated true correlation between *g* loadings and group differences.

Group Identity as Moderator

The final estimated true correlation of .71 could be moderated by the groups that have been compared in our analysis of all IQ batteries. Different groups (Blacks, Whites, Native-Americans, Hispanics) within the US have been compared and groups from Europe and Africa have been included in the meta-analysis. To test whether the specific group identity moderates the relation between group differences and *g,* first, the dataset has been split up into four subsets. This resulted in 1) a subset of studies on Black/White differences, 2) a subset of studies on White/Hispanic differences, 3) a cluster of studies on Dutch/immigrant differences, and 4) a subset of studies on group differences in Africa. This last subset contained only three studies, so this cluster was not used for the moderator analyses (see Table 6). Additionally, the Black/White and the White/Hispanic subsets were combined to see the results for all North-American studies. The results from these moderator analyses are compared with the results of the outcomes of the psychometric meta-analysis of all IQ batteries minus six outliers. That is, the values of rho-4 and the percentage of variance from the moderator analyses are compared to the dataset of Spearman's hypothesis on IQ batteries where outliers have been left out of the analysis. For the Black/White cluster the same outliers that have been excluded from the initial number of 38 data points were also excluded in these moderator analyses. This resulted in a Black/White cluster of 11 data points (excluding Department of Defense,1982; and Jensen & Faulstich, 1988), and a Hispanic/White cluster consisting of four data points, which included the study by Valencia and Rankin(1986). The immigrant/Dutch cluster consisted of all the studies conducted in the Netherlands. It could be that clustering the subsets would moderate the data resulting in a value of the percentage of variance explained that is closer to the theoretically optimal value of 100. It may also be the case that the values of rho-4 differ substantially by group. The results were that in the first dataset on Black and White groups in the US rho-4 decreases from .66 to .63. In the second cluster of Hispanics/Whites rho-4 increases from .66 to .75. In the last subset of studies containing immigrant and Dutch groups rho-4 decreases to .63, but with a substantially lower percentage of variance explained by the four artifacts: 38. These results show that the value of rho-4 is quite similar in the various groups and thereby disconfirm the hypothesis that different group identity used in the overall meta-analysis act as a moderator.

Test battery as a moderator

Because of the substantially lower outcomes of the RAKIT within the Dutch cluster of studies on immigrants and Dutch we hypothesized that there was a another specific effect operating for this group of studies, namely type of test battery. We hypothesized that when a division would be made within the Dutch studies between the Dutch RAKIT studies and the Dutch studies where another test battery had been used it would result in a larger percentage of variance explained within each group. The collection called ‘Dutch other studies’ includes all studies using any IQ battery other than the RAKIT. The study by Helms-Lorenz et al. (2003) is not entered in one of the two separate analyses because one of the two IQ batteries used was the RAKIT, and so the results could have been contaminated. The results from this moderator analysis are compared with the results of the outcome of the psychometric meta-analysis of all Dutch studies. In the group of RAKIT studies rho-4 remains the same, the *SD* of rho decreased from .19 to .13, and the percentage variance explained increased from 38 to 64. When one outlier is removed, namely Tolboom’s (2000) data on 5.8-year-old Dutch and Moroccans, the amount of variance explained increased dramatically to 119. In the ‘Dutch other studies’ cluster rho-4 is much higher (.72) than for Dutch all studies (.63), but the *SD* of rho is much lower, namely .07. However, the percentage of variance explained is still low, namely 48. When removing one outlier, namely the Surinamese group from te Nijenhuis and van der Flier (1997), rho-4 increased even more, namely a value of .79. The *SD* of rho decreased further to .04 and the percentage of variance explained increased drastically to 164.

Dutch GATB as a standard for restriction of range

Table 7 presents the results of the psychometric meta-analysis of the thirty-eight data points where the Dutch GATB has been used as the standard for the correction for restriction of range. It has the same format as Table 6. The analysis of all 38 data points yields an estimated correlation (rho-4) of .73, with only 18% of the variance in the observed correlations explained by artifactual errors. The same extreme outliers were left out of this analysis. After first leaving out three extreme outliers, the correlation (rho-4) increased to a value of .84, the *SD* of rho-4 decreased drastically, and the amount of variance explained in rho-4 by artifacts increased to 58 %. We also checked what would happen when instead of three, six outliers were removed. The value of rho-4 did not change drastically, but the *SD* of rho-4 decreased drastically to a value of .01, and the percentage of variance explained by artifacts increased to 98. Finally, a correction for deviation from perfect construct validity in *g* took place, using the value of 7.5 %. This resulted in a value of .91 for the final estimated true correlation between *g* loadings and group differences.

Group Identity as Moderator

Although artifacts explained 98 % of the variance in the data points, the final estimated true correlation of .91 could in theory be moderated by the groups that have been compared in our analysis of all IQ batteries. Groups were formed in the same way and the same outliers were excluded as in the previous set of analyses. In the first dataset on Black and White groups in the US rho-4 remains at a value of .85. In the second cluster of Hispanics/Whites rho-4 increases from .85 to .89. In the last group of studies containing Dutch and immigrant groups rho-4 decreases to .78, but with a substantial lower percentage of variance explained by the four artifacts: 30 %. These results disconfirm the hypothesis that different groups used in the overall meta-analysis act as a moderator: the values of rho-4 in the three groups are highly similar.

Test battery as a moderator

The composition of the groups within this analysis are composed is the same as in the previous set of analyses. In the group of RAKIT studies rho-4 decreased to .74, the *SD* of rho decreased from .15 to .12, and the variance explained increased from 30 % to 66 %. When one outlier is removed, namely Tolboom's (2000) 5.8-year-old Dutch and Moroccans, the amount of variance explained increased dramatically to 123 %. In the Dutch other studies cluster rho-4 is much higher (.87) than for Dutch all studies (.78). The *SD* of rho is much lower, namely .03. Moreover, the variance explained increased as well, namely to 70 %. When removing one outlier, namely te Nijenhuis and van der Flier's (1997) Surinamese group, rho-4 increased even more, namely to .91. The *SD* of rho decreased further to 0 and the variance explained increased drastically to 288%.

Tables 6 and 7 show percentages of variance explained as being larger than 100 % . This phenomenon is called "second-order sampling error", and results from the sampling of studies in a meta-analysis. Percentages of variance explained greater than 100% are not uncommon when only a limited number of studies are included in an analysis. The proper conclusion is that all the variance is explained by statistical artifacts (see Hunter & Schmidt, 2004, pp. 399-401, for an extensive discussion).

Table 5

Studies of Correlations Between g Loadings and Group Differences

references by cluster	test battery	*r*	N_{harmonic}	mean age (range)
Black/White differences				
Jensen & Reynolds (1982) [1]	WISC-R	.77	524	11.5 (6.5-16.5)
Mercer (1984) [1]	WISC-R	.71	643	8 [2] (5-11)
Nichols (1972) [1]	WISC-R	.75	1,666	7 (7)
Department of Defense (1982) [1]	ASVAB	.30	3,247	19.5 [2] (16-23)
Kaufman & Kaufman (1983) [1]	K-ABC	.60	608	7.5 [2] (2.6-12.5)
Kane & Brand (2008)	WISC-III	.58	545	11 [2] (6-16)
Naglieri & Jensen (1987)	WISC-R, K-ABC	.71	86	10.7 (9.4-12.4)
Jensen & Faulstich (1988)	WAIS-R	-.03	120	28 (18-70)
Department of Labor [1]	GATB Aptitudes	.71	3,120	40 (16-70)
Hennessy & Merrifield (1976) [1]	CGP	.66	697	18 (17-19)
National Longitudinal Study [1]	CGP, SAT, ACT	.78	3,348	18 (16-23)
Nyborg & Jensen (2000)	Various tests	.81	879	19.9 (17-25)
Montie & Fagan (1988)	Stanford Binet 3 (LM)	.30	86	3 (3.0-3.9)
Hispanic/White differences				
Hartmann, Kruuse, & Nyborg (2007)	Various tests	.71	345	19.9 (17-25)
Hartmann, Kruuse, & Nyborg (2007)	ASVAB	.74	2,737	19.6 (15-24)
Reynolds, Willson, & Ramsey (1999)	WISC-R	W/H = .84	238	10.12 (6-16)
Valencia & Rankin (1986)	K-ABC	W/H = .15 (.79) [4]	100	11 (10-12.5)
Native-American/White differences				
Reynolds, Willson, & Ramsey (1999)[5]	WISC-R	W/NA = .74	238	10.12 (6-16)
Dutch/Immigrant differences				
Te Nijenhuis, Tolboom, Resing, & Bleichrodt (2004) 7.8 years	RAKIT	D/S&NA = .56 (-.12)[4] D/T = .55 (-.03)[4] D/M = .39 (.14)[4]	93 104 92	7.8 (7.6-7.10)
Tolboom (2000) 5.8 years	RAKIT	D/S&NA = .45 (.27)[4] D/T = .46 (.004)[4] D/M= .09 (.03)[4]	93 94 95	5.8 (5.6-5.10)
Tolboom (2000) 9.8 years	RAKIT	D/S&NA = .82 (.59)[4] D/T = .43 (.02)[4] D/M= .42 (.22)[4]	95 95 93	9.8 (9.6-9.10)
Helms-Lorenz, Van de Vijver, & Poortinga (2003)	RAKIT & SON-R	.70 (.10)[4]	580	9 [2] (6-12)
Te Nijenhuis & Van der Flier (1997)	GATB	D/S = .72 D/Ant = .80 (.77)[4] D/N.A.= .87 (.84) [4] D/T = .82 (.70)[4]	643 218 277 410	28 [2] (18-55)
Te Nijenhuis & Van der Flier (2005)	GATB	.77 (.76)[4]	78	28.7 (20-40) [3]
Te Nijenhuis, Evers, & Mur (2000)	DAT	.84 (.76)[4]	165	12 (12-13)
African samples				

Rushton (2001)	WISC-R	US/African = .71	174	14 (13-15)
Rushton & Jensen (2003)	WISC-R	Z/W(UK) = .36	203	13 (12-14)
Lynn & Owen (1994)	JAT	W/I = .21 (.14)[4] W/B = -.175 (.14)[4]	1,070 1,070	15.5 (15-16)

Note. [1] These studies were taken from Jensen (1985). [2] Mean age was estimated based on the age range reported in the study. [3] Age range was estimated based on the mean age reported in the study. [4] Values in brackets are correlations before language biased subtests had been removed. [5] This study did not fit in a group so it was left out of the moderator analyses. W = Whites; H = Hispanics; NA = Native-Americans; D = Dutch; S&NA = Surinamese and Netherlands-Antilleans; T = Turks; M = Moroccans; S = Surinamese; Ant. = Antilleans; N.A.= North-Africans; Z = Zimbabwean; I = Indians; B = Blacks.

Table 6

Meta-analytical Results for Correlation Between Group Differences and g Loadings After Corrections for Reliability, Restriction of Range, and Imperfect Construct Validity; For Restriction of Range the Wechsler is Taken as Reference

predictor	*K*	*N*	*r*	SD_r	rho-4	$SD_{\text{rho-4}}$	rho-5	%VE	80% CI
group differences[1]	38	24,969	.61	.22	.57	.23	.61	24 %	.27-.86
group differences minus 3 outliers	35	19,582	.71	.12	.65	.07	.70	77 %	.56-.74
group differences minus 6 outliers	32	19,267	.72	.09	.66	0	.71	130 %	.66-.66
moderator									
Black/White differences + Hispanic/White differences	17	18,989	.65	.18	.60	.18	.65	28 %	.36-.83
Black/White differences + Hispanic/White differences (minus 1 outlier)	16	15,742	.72	.10	.65	.03	.70	93 %	.60-.69
Black/White differences	13	15,569	.63	.19	.57	.20	.61	23 %	.32-.82
Black/White differences (minus 1 outlier)	12	12,322	.72	.10	.63	.05	.68	86 %	.57-.69
Black/White differences (minus 2 outliers)	11	12,202	.73	.07	.63	0	.68	196 %	.63-.63
Hispanic/White differences	4	3,420	.73	.10	.75	0	.81	100 %	.75-.75
Dutch all studies	16	3,225	.69	.17	.63	.19	.68	38 %	.39-.87
Dutch RAKIT studies	9	854	.46	.16	.62	.13	.67	64 %	.46-.79
Dutch RAKIT studies (minus 1 outlier)	8	759	.51	.11	.67	.14	.72	119 %	.67-.67
Dutch other studies	6	1,791	.79	.05	.72	.07	.77	48 %	.63-.80
Dutch other studies (minus 1 outlier)	5	1,148	.83	.02	.79	.04	.85	164 %	.79-.79

Note. [1] Meta-analytical results for correlations between *g* loadings and *d* (group differences). *K* = number of correlations; *N* = total harmonic mean; *r* = mean observed correlation (sample-size weighted); SD_r = standard deviation of observed correlation; rho-4 = observed correlation corrected for sampling error, unreliability, and range restriction; $SD_{\text{rho-4}}$ = standard deviation of correlation; rho-5= true correlation (observed correlation corrected for sampling error, unreliability, range restriction, and imperfect construct validity); %VE = percentage of variance accounted for by artifactual errors; 80% CI = 80% credibility interval.

Table 7

Meta-analytical Results for Correlation Between Group Differences and g Loadings After Corrections for Reliability, Restriction of Range, and Imperfect Construct Validity; For Restriction of Range the Dutch GATB is Taken as Reference

predictor	*K*	*N*	*r*	SD_r	rho-4	$SD_{\text{rho-4}}$	rho-5	%VE	80% CI
group differences[1]	38	24,969	.61	.22	.73	.20	.78	18 %	.48-.98
group differences minus 3 outliers	35	19,582	.71	.12	.84	.07	.90	58 %	.75-.92
group differences minus 6 outliers	32	19,267	.72	.09	.85	.01	.91	98 %	.83-.86
moderator									
Black/White differences + Hispanic/White differences	17	18,989	.65	.18	.78	.15	.84	21 %	.60-.97
Black/White differences + Hispanic/White and differences (minus 1 outlier)	16	15,742	.72	.10	.85	.05	.91	65 %	.79-.91
Black/White differences	13	15,569	.63	.19	.76	.16	.82	15 %	.55-.97
Black/White differences (minus 1 outlier)	12	12,322	.72	.10	.84	.06	.90	50 %	.76-.92
Black/White differences (minus 2 outlier)	11	12,202	.73	.07	.85	0	.91	108 %	.85-.85
Hispanic/White differences	4	3,420	.73	.10	.89	.01	.96	98 %	.88-.90
Dutch all studies	16	3,225	.69	.17	.78	.15	.84	30 %	.58-.97
Dutch RAKIT studies	9	854	.46	.16	.74	.12	.80	66 %	.59-.90
Dutch RAKIT studies (minus 1 outlier)	8	759	.51	.11	.79	0	.85	123 %	.79-.79
Dutch other studies	6	1,791	.79	.05	.87	.03	.94	70 %	.82-.91
Dutch other studies (minus 1 outlier)	5	1,148	.83	.02	.91	0	.98	288 %	.91-.91

Note. [1] Meta-analytical results for correlations between *g* loadings and *d* (group differences). *K* = number of correlations; *N* = total harmonic mean; *r* = mean observed correlation (sample-size weighted); SD_r = standard deviation of observed correlation; rho-4 = observed correlation corrected for sampling error, unreliability, and range restriction; $SD_{\text{rho-4}}$ = standard deviation of correlation; rho-5= true correlation (observed correlation corrected for sampling error, unreliability, range restriction, and imperfect construct validity); %VE = percentage of variance accounted for by artifactual errors; 80% CI = 80% credibility interval.

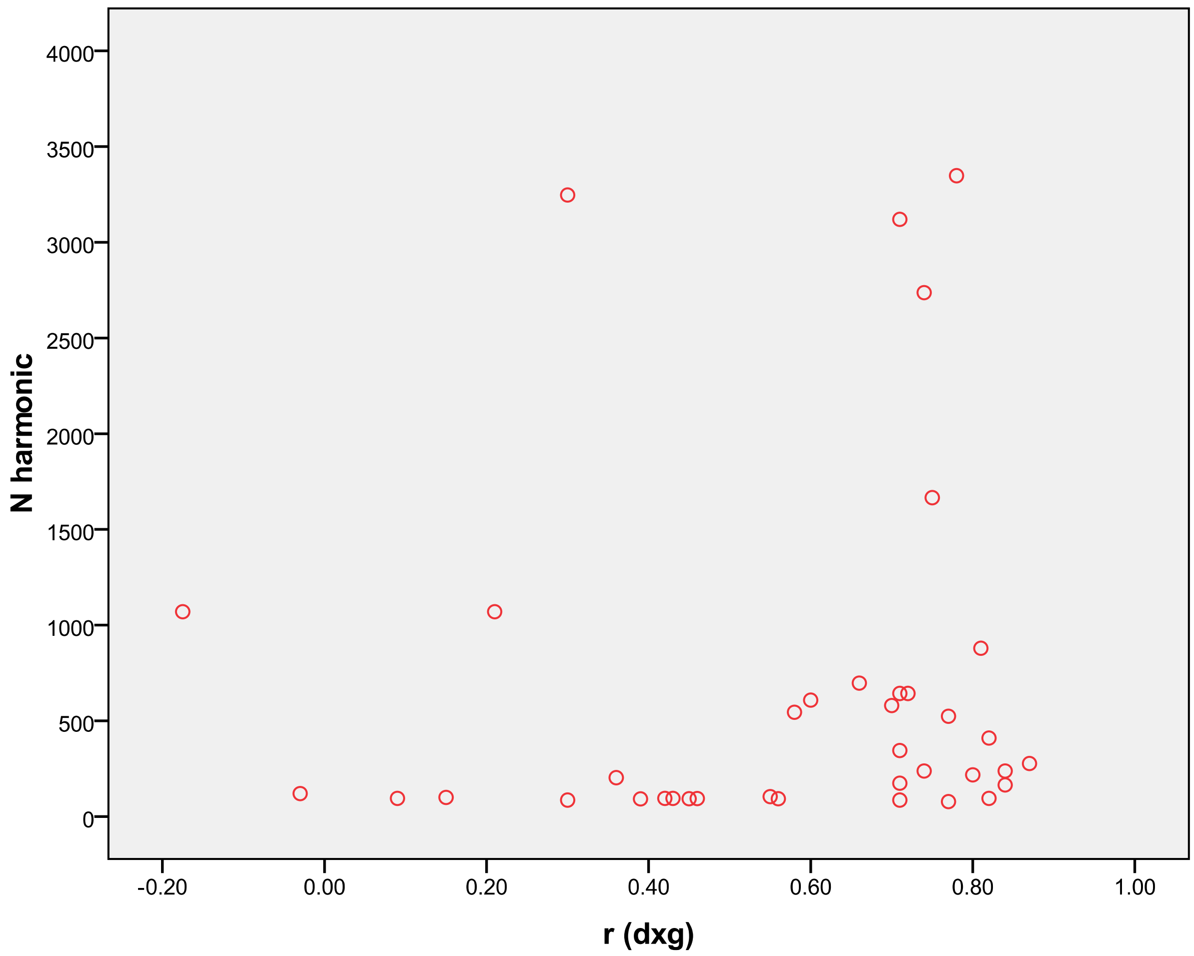

Figure 12

Scatter Plot of Correlations (*d* x *g*) *and Harmonic Mean for all Studies*

Bare-bones analysis of educational and training criteria

A bare-bones psychometric meta-analysis was carried out to estimate the size of the correlation between group differences in educational and training criteria and *g* loadings. A bare-bones psychometric meta-analysis estimates how much of the observed variance in findings across studies is due to sample size alone. Criteria with a substantial language bias were removed from the analysis in a similar manner as with the IQ batteries; an example is given in Figure 13 and Figure 14. The results of the studies on the correlation between *g* loadings and score differences in educational and training criteria are shown in Table 8. Mostly high correlations are found between *g* loadings and differences in educational and training criteria.

Table 9 shows that the bare-bones meta-analysis yields a correlation between group differences in educational and training criteria and *g* loadings of .67, using only one simple correction for sample size. The *SD* of *r* is large, namely a value of .27. Moreover, the variance accounted for by artifactual errors is a negligible at 4 %. When two statistical outliers are removed, namely Tolboom's (2000) data on 9.8-year-old Turks and Moroccans, the correlation increased to a value of .79 but the variance explained is still low, namely 17 %. Most likely this is a result of the fact that all groups have a highly comparable sample size, meaning there is a severe restriction of range in *N*.

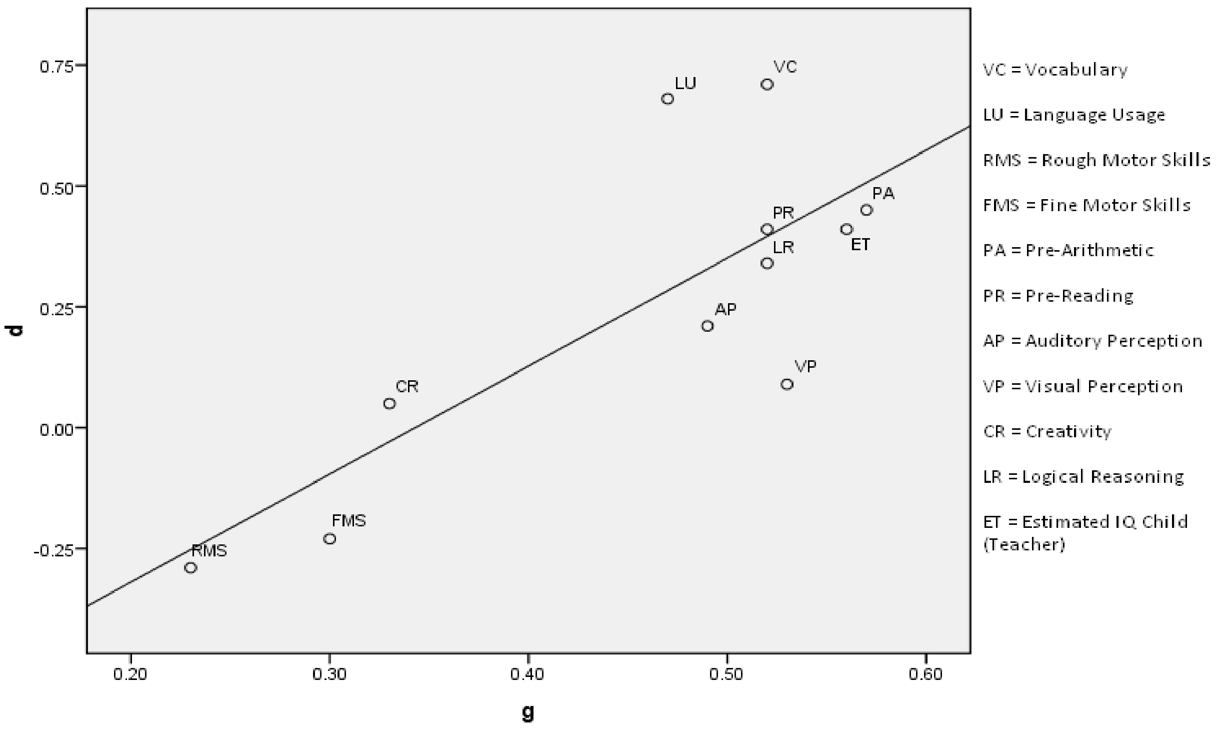

Figure 13

Scatter Plot of School Criteria by g: Differences Between Scores of Dutch and Moroccan Children of 5.8 Years Including Language Biased School Criteria

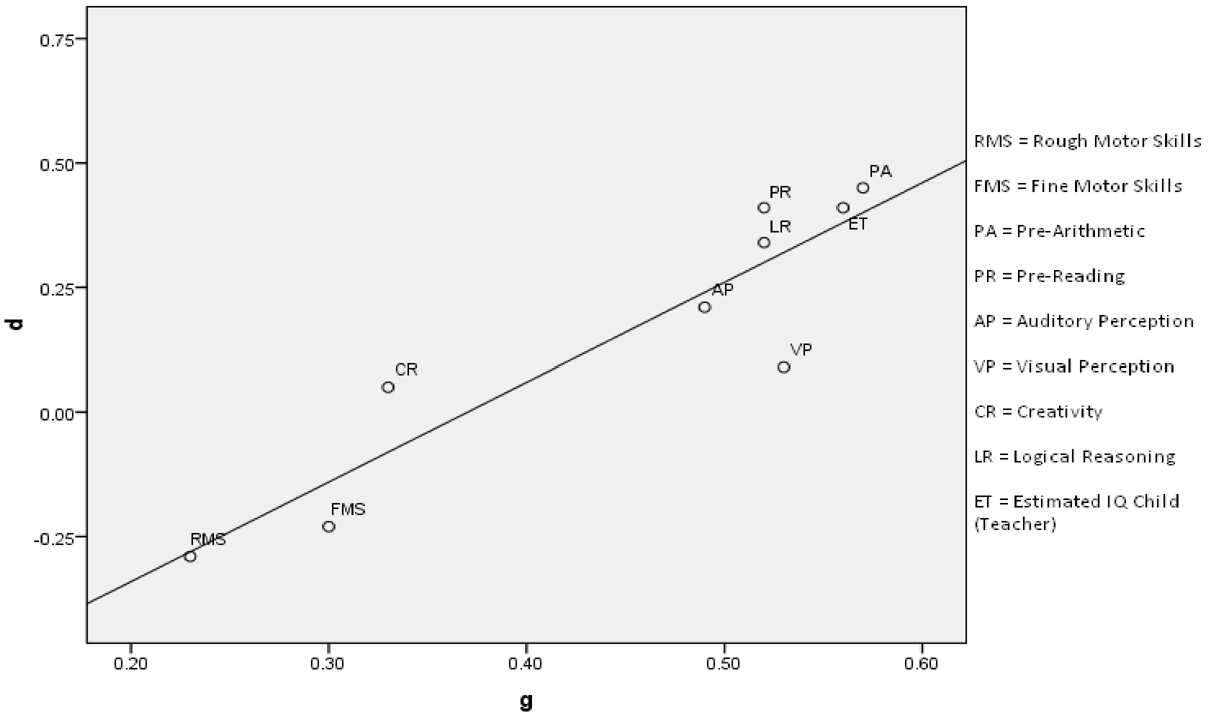

Figure 14

Scatter Plot of School Criteria by g; Differences Between Scores of Dutch and Moroccan Children of 5.8 Years Excluding Language Biased School Criteria

Table 8

Studies of Correlations Between g Loadings and Differences in Educational and Training Criteria Between Dutch Groups and Several Immigrant Groups

study	criteria	groups compared	*r*	N_{harmonic}	mean age (range)
Te Nijenhuis & Van der Flier (2005)	training outcomes	D/Imm.	.66 (.47)[1]	78	28.7 (20-40)
Te Nijenhuis, Tolboom, Resing, & Bleichrodt (2004) 7.8 years	school criteria	D/S&NA D/T D/M	.71 (.68)[1] .69 (.57)[1] .94 (.81)[1]	93 104 92	7.8 (7.6-7.10)
Tolboom (2000) 5.8 years	school criteria	D/S&NA D/T D/M	.86 (.86)[1] .79 (.65)[1] .92 (.79)[1]	93 94 95	5.8 (5.6-5.10)
Tolboom (2000) 9.8 years	school criteria	D/S&NA D/T D/M	.72 [2] -.08 [2] .30 [2]	95 95 93	9.8 (9.6-9.10)
Te Nijenhuis, Evers, & Mur (2000)	school criteria	D/Imm.	.80 (.75) [1]	165	12 (12-13)

Note. [1] Values in brackets are correlations before language biased criteria had been removed. [2] In these samples no language-biased criteria have been identified. D = Dutch; Imm. = Immigrants; S&NA = Surinamese and Netherlands-Antilleans; T = Turks; M = Moroccans.

Table 9

Bare-bones Meta-analytical Results for Correlations Between g Loadings and Educational and Training Criteria

topic	*K*	*N*	*r*	SD_r	%VE
Educational and training criteria	11	1,093	.67	.27	4 %
Educational and training criteria minus 2 outliers	9	909	.79	.08	17 %

Table 10

Outcomes of Spearman's Hypothesis in Studies Using Raven's Progressive Matrices at the Item Level

study	groups compared	*r*
Rushton & Skuy (2000)	White and African	.39
Rushton, Skuy, & Fridjhon (2002)	White, Indian, and African	W/I = .15 W/A = .27
Rushton, Skuy, & Fridjhon (2003)	White, Indian, and African	W/I = .27 W/A = .38
Rushton (2002)	White, Indian, and Colored	W/I = .36[1] W/C = .57[1]
Rushton, Cvorovic, & Bons, (2007)	Majority Serbians and Roma	.54

Note. [1] Correlations are Spearman's rank-order correlations (rho's).

Conclusion and Discussion

The true correlation between group differences in mean intelligence test score and *g* loadings is strong, as predicted by Spearnan's hypothesis. The true correlation is independent of the groups that are compared and independent of the country in which the study was conducted.

In this study two different reference tests have been used for the restriction of range in *g* loadings. These two methods do not produce similar outcomes: the outcomes of the method where the Dutch GATB is used as the reference for the restriction of range correction are clearly higher in terms of the value of rho-4. In the general discussion the best option for the restriction of range correction is discussed. Furthermore, our bare-bones meta-analysis showed that for educational and training outcomes the true estimated correlation between *d* and *g* is even higher, namely $r = .79$. The low percentage of variance explained by sample size is most likely due to the highly comparable sample sizes used in the studies that have been analyzed. The outcomes of the studies using Raven's Progressive Matrices are rather low compared to the results of for example the studies of intelligence batteries. A highly plausible explanation for this finding is the unreliability of testing on the item level; strong corrections for unreliability would increase the values of the correlations and would bring them closer to the rho-5's reported for test batteries.

General Discussion

Intelligence tests are the best predictors of job performance and educational performance and many studies have been conducted on differences in mean intelligence test scores between groups. Often cultural factors are used to explain these differences. However, since Charles Spearman suggested back in 1927 that the different magnitude of the mean Black/White difference on various subtests of an IQ battery is a function of each test's complexity numerous studies have investigated this so called Spearman's hypothesis. In this study we collected the complete empirical literature and conducted a meta-analysis. The findings clearly show that the true correlation between mean group differences and *g* loadings is strong: a correlation of .71 based on the Wechsler tests as a reference for the restriction of range correction and a correlation of .91 when the Dutch GATB was taken as a reference. Probably the GATB is a better reference, as its variance in *g* loadings is closer to the variance in *g* loadings from a theoretically optimal test battery, measuring all broad abilities of Carroll's (1993) model. Also, the correlations between group differences and *g* loadings do not differ by group; some outcomes are even virtually identical.

Our study on language bias shows that language-biased subtests underestimate IQ for minority groups, but virtually all effects are small. Furthermore, our study on the correction value for imperfectly measuring the construct of *g* estimated it to be 7.5 %, which is substantially smaller than the value used in previous research.

Based on our large-scale meta-analysis it is highly plausible that all group differences on subtests of IQ batteries can be explained using: (1) *g* loadings; (2) language bias; and (3) group differences on a small number of broad abilities such as short-term memory and broad visual perception. This is another nail in the coffin of the explanation of group differences in mean IQ scores being the result of cultural factors because there appears to be simply no variance left for these alleged cultural factors to explain.

In Study 1 we focused on the question of the degree to which language bias causes a disadvantage for minority groups and how large an underestimation of IQ it produces. Some of the samples used in our analyses consisted of test takers for whom the language of the test was not their first language or who were non-native speakers of the language of the test. We found that in virtually all cases this led to only small underestimates of their IQ with a maximum of three IQ points. The only exception was the K-ABC, with no less than five of the thirteen subtests susceptible to language bias, which resulted in an underestimate of ten IQ points, a large effect. Of the Dutch batteries the RAKIT, with four of the twelve subtests susceptible to language bias, was the most language-biased battery. This battery underestimates IQ with an average of 3 IQ points; of course this is three points too much, but it is still a small effect. A recent study by Wicherts (2007) claims an underestimate of no less than seven IQ points for the Dutch RAKIT. The discrepancy between our meta-analytical findings and Wicherts' findings based on just one data set is so large, that one might ask the question whether there is a fundamental flaw in Wicherts' analysis. Moreover, the contrast is even larger with the outcomes for the GATB, which shows an underestimate of IQ of about one IQ point, which is very small.

The value for the fifth correction for artifacts was investigated in the second Study. In previous research a correction factor of 10 percent was used, but our new analysis leads to an improved value for the correction factor of 7.5 %, which is substantially lower. In previous research a conservative value was taken but this correction factor leads to a substantial overestimate in comparison with the new value. In this study the correction factor was applied to all collections of studies, whereas in the future values for specific collections of studies can be computed to get more precise outcomes.

A psychometric meta-analysis of Spearman's hypothesis was the focus of the third Study. There are two different forms of Spearman's hypothesis: a strong and a weak form. The strong form states that all mean differences between groups can be explained by *g* loadings, the weak form states that the differences are predominantly in *g*. Jensen (1998) has shown that there are group differences besides *g*, namely on broad and narrow abilities. Tests that show consistently larger Black/White differences than are expected based on their *g* loadings loaded on a spatial visualization factor. Tests that show a smaller Black/White differences than is predicted by the tests' *g* loading are those that are loaded on a short term memory (STM) factor. The fact that tests that are heavily loaded on either the spatial factor or STM factors consistently cause small deviations from the result predicted by the strong form of Spearman's hypothesis dictates that this form must be rejected. The weak form of Spearman's hypothesis, however, is strongly confirmed. The true correlation between mean group differences and *g* loadings is strong: a of .91 with the Dutch GATB taken as a reference for the corrections for restriction of range in *g* loadings and virtually all the variance in the data points explained by five statistical artifacts.

Spearman's hypothesis has been used to test for bias in tests, and the higher the correlation between *d* and *g*, the less need there is for explanations involving cultural bias. When considering group differences on IQ tests, taking into account (1) the effect of language bias, and (2) well-known group differences on the broad abilities Short Term Memory and Broad Visual Perception, it appears that there is virtually no room for explanations based on cultural bias. Our re-analysis of the high-quality, highly-innovative paper by Helms-Lorenz et al. (2004) clearly shows their data set is non-optimal. These researchers used an unrepresentative set of tests: (1) a large part of the subtests administered measured Broad Visual Perception, and (2) three subtests show a substantial language bias. It strongly appears that the main test battery used by these researchers yields data points that are outliers in the meta-analysis.

It seems highly plausible that when taking into account group differences on *g,* short-term memory, spatial ability, and proficiency in the language of the test group differences on all subtests of a battery can be near perfectly explained. This position is supported by outcomes of the analyses on the educational and training data, which show very high correlations between *d* and *g* without any correction for unreliability and restriction of range. One could argue that the findings for the Raven's are also in support. The correlations between group differences and complexity at the item level of the Raven's are substantial but also much lower than the sizes of the correlations found for the IQ batteries. However, item scores are highly unreliable and correcting for this strong unreliability might lead to values of rho that are similar to the values of rho (the true correlation resulting from psychometric meta-analysis) found for test batteries.

Recent psychometric meta-analyses have clearly shown that *g* loadings correlate highly with measures of heritability. te Nijenhuis and Grimen (2007) show that *g* loadings of subtests correlate perfectly with these subtests' heritability coefficients. Moreover, te Nijenhuis and Franssen (2010) show that inbreeding depression correlates .85 with *g* loadings. This strongly suggests that *g* loadings and heritability coefficients may be interchangeable. This in turn suggests that the high correlation between *g* loadings and group differences could imply that mean group differences have a substantial genetic component. However, this is not necessarily the case, as the score patterns of biological factors, such as better nutrition and better health care for pregnant women, may mimic the score pattern of the heritability coefficient. At the present, these effects are impossible to disentangle, as all the available research is correlational and not experimental.

Further, a strong finding from the present analysis is that the results from US and Europe, though looking at very different populations, yield highly comparable outcomes. This picture emerged from the moderator analyses where a number of outcomes were highly similar and some were virtually identical. This contributes to the clearer understanding of group differences on a global scale. What has to be noted in this respect, however, is that the vast majority of studies of Spearman's hypothesis have been conducted with Blacks and Whites in the US. A smaller cluster of studies has been conducted in the Netherlands. We recommend conducting additional research in parts of the world that have not yet been studied. For instance, although North-East Asians are known to have a higher mean IQ than Europeans and US Whites, there are no studies testing Spearman's hypothesis using intelligence batteries on North-East Asian samples.

Practical implications

IQ tests are important instruments for selection and placement in work and educational settings, and there are large differences in mean IQ scores between groups. The present study makes a strong empirical contribution to the important discussion about the nature of group differences in intelligence. It strongly suggests that IQ batteries are not culturally biased, apart from a small effect for language bias when there is a substantial number of subtest with a strong verbal component in one or more specific tests. This provides strong support for the validity of these tests because group differences in intelligence are not attributable to cultural bias. It appears that with the exception of language-biased subtests, one can confidently use IQ batteries in work and educational settings.

Limitations of the study

In this meta-analysis our initial data set of data points comprised thirty-eight correlations. We choose to exclude no less than six data points, and this choice was founded on statistical and theoretical considerations. Although the database for the meta-analysis is huge, it would seem that even more data points are needed to create a more solid meta-analytical database. For example, the Dutch database is too one-sided in the studies that have been included: ninety percent of the studies use either the RAKIT or the GATB. It would be a good idea to carry out some additional basic studies on Spearman's hypothesis employing other test batteries. Moreover, in this way it could also be investigated whether clustering the different immigrant groups in the analysis yields more reliable results. Analyses where separate immigrant groups have been compared against a Dutch majority group repeatedly show lower outcomes than a combination of the different immigrant groups. Apart from the Black/White and the Dutch studies the number of data points within the other clusters was small. For example, the Hispanic cluster consisted only of four studies and for both Native-Americans and Gypsies there were only one study each. The number of studies of these groups should also be increased.

Methodological refinements

In this study the debate is open as to which is the best reference group for the correction of restriction of range in *g* loadings. The most commonly used intelligence battery in the world, the Wechsler, was compared against the battery with the largest standard deviation in *g* loadings in our sample, the Dutch GATB. At this moment we could not make a definite choice which of the two batteries is the better. In theory, Carroll's model (1993) could be used as a basis for constructing a theoretically perfect IQ battery. This would mean that for every one of Carroll's twelve broad abilities three representative subtests would have to be collected. The variance in *g* loadings would then be taken as the standard. Another option would be to use the *SD* of the *g* loadings of the Woodcock-Johnson test battery, which is explicitly modeled on the Carroll model. As stated before, the *SD* of *g* loadings of these batteries will most likely be closer to the *SD* of the *g* loadings of the GATB than the *SD* of the *g* loadings of the Wechsler.

Conclusion

Mean group differences in scores on cognitive-loaded instruments are well documented over time and around the world. A meta-analytic test of Spearman's hypothesis was carried out. Mean differences in intelligence between groups can be largely explained by cognitive complexity and the present study shows clearly that there is simply no support for cultural bias as an explanation of these group differences. Comparing groups, whether in the US or in Europe, produced highly similar outcomes.

References

References marked with an asterisk indicate studies included in the meta-analyses and in the exploratory studies.

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

van den Berg, M. (2001). Culture and language in transition : A festschrift in honor of H.L. Cox's 65th birthday. *Volkskunde, 102,* 86-92.

Bleichrodt, N., Resing, W. C. M., Drenth, P. J. D., & Zaal, J. N. (1987). *Intelligentiemeting bij kinderen* [The measurement of children's intelligence]. Lisse, The Netherlands: Swets.

Carretta, T. R. (1997). Group differences on US Air Force Pilot Selection Tests. *Sex and Ethnic Group Differences, 5,* 115-127.

Carretta, T. R., & Ree, M. J. (1995). Air Force Officer Qualifying Test validity for predicting pilot training performance. *Journal of Business and Psychology, 9,* 379-388.

Carretta, T. R., & Ree, M. J. (1995). Near identity of cognitive structure in sex and ethnic groups. *Personality and Individual Differences, 19,* 149-155.

Carroll, J. B. (1993). *Human Cognitive Abilities*. Cambridge University Press: Cambridge.

Cascio, W. F. (1991). *Applied psychology in personnel management* (4th ed.). Englewood Cliffs, NJ: Prentice Hall.

Centers for Disease Control (1988). Health status of Vietnam veterans. *Journal of the American Medical Association*, *259*, 2701-719.

Colom, R., Juan-Espinosa, M., Abad, F., & Garcia, L. F. (2000). Negligible sex differences in general intelligence. *Intelligence, 28,* 57-68.

Evers, A., te Nijenhuis, J., & van der Flier, H. (2005). Ethnic bias and fairness in personnel selection: Evidence and consequences. In A. Evers, N. Anderson, & O. Voskuijl (Eds.), *The Blackwell handbook of personnel selection* (pp. 306-328). Oxford: Blackwell.

* Goldstein, H. W., Yusko, K. P., Braverman, E. P., Smith, D. B., & Chung, B. (1998). The role of cognitive ability in the subgroup differences and incremental validity of assessment center exercises. *Personnel Psychology*, *51*(2), 357-374.

* Goldstein, H. W., Yusko, K. P., & Nicolopoulos, V. (2001). Exploring Black-White subgroup differences of managerial competencies, *Personnel Psychology*, *54,* 783-807.

Gottfredson, L. S. (1997). Why *g* matters: The complexity of everyday life. *Intelligence, 24*, 79-132.

van Haasen, P. P., de Bruyn, E. E. J., Pijl, Y. J., Poortinga, Y. H., Lutje-Spelberg, H. C., et al. (1986). *Wechsler Intelligence Scale for Children-Revised, Dutch Version.* Lisse: Swets and Zetlinger B.V.

* Hartmann, P., Kruuse, N. H. S., Nyborg, H. (2007). Testing the cross-racial generality of Spearman's hypothesis in two samples. *Intelligence, 35,* 47-57.

Heneman, H. G., & Heneman, R. L. (1994). *Staffing Organizations.* Middleton, WI: Mendota House.

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

Helms-Lorenz, M. (2001). *Assessing cultural influences on cognitive test performance: a study with migrant children in the Netherlands.* Tilburg University.

* Hennessy, J. J., & Merrifield, P. R. (1976). A comparison of the factor structures of mental abilities in four ethnic groups. *Journal of Educational Psycholog, 68,* 754-759.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis*. London: Sage.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis (2nd ed.).* London: Sage.

Ja-Song, M., & Lynn, R. (1992). Reaction times and intelligence in Korean children. *Journal of Psychology, 126,* 421–428.

Jensen, A. R. (1980). *Bias in Mental Testing*. London: Methuen.

Jensen, A. R. (1985). The nature of the Black-White difference on various psychometric tests: Spearman's hypothesis. *The Behavioral and Brain Sciences, 8,* 193-263.

Jensen, A. R. (1993). Spearman's hypothesis tested with chronometric information processing tasks. *Intelligence, 17,* 47-77.

Jensen, A. R. (1998). *The g factor: The science of mental ability*. London: Praeger.

* Jensen, A. R. & Faulstich, M. E. (1988). Difference between prisoners and the genral population in psychometric *g. Personality and Individual Differences, 9,* 925-928.

* Jensen, A. R., & Reynolds, C. R. (1982). Race, social class and ability patterns on the WISC-R. *Personality and Individual Differences, 3,* 423-438.

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

Jensen, A. R., & Whang, P. A. (1993). Reaction-times and intelligence: A comparison of Chinese-American and Anglo-American children. *Journal of Biosocial Science, 23,* 397-410.

Jensen, A. R., & Whang, P. A. (1994). Speed of accessing arithemtic facts in long-term memory: A comparison of Chinese-American and Anglo-American children. *Contemporary Educational Psychology, 19,* 1-12.

* Kane, H., & Brand, C. (2008). Spearman's hypothesis: Support from the Wechsler Intelligence Scale for children,third edition. *Mankind Quarterly, 49,* 3-22.

* Kaufman, A. S., & Kaufman, N. L. (1983). Kaufman Assessment Battery for Children: Interpretive manual. American Guidance Service.

Klockars, A.J., & Sax, G. (1987). *Multiple comparisons.* Newbury Park, CL: Sage.

Lopez, E. C. (1997). The cognitive assessment of limited English proficient and bilingual children. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), Contemporary *intellectual assessment: theories, tests, and issues* (pp. 503–516). New York: The Guilford Press.

Lynn, R., Chan, J. W. C., & Eysenck, H. J. (1991). Reaction times and intelligence in Chinese British children. *Perceptual and Motor Skills, 72,* 443–452.

Lynn, R., & Holmshaw, M. (1990). Black-White differences in reaction times and intelligence. *Social Behavior and Personality, 18,* 299-308.

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

Lynn, R., & Shigehisa, T. (1991). Reaction times and intelligence: A comparison of Japanese and British children. *Journal of Biosocial Science, 23,* 409-416.

* McDaniel, M. A., Hartman, N. S., Whetzel, D. L, & Grubb, W. L. (2007). Situational judgment tests, response instructions, and validity: A meta-analysis. *Personnel Psychology, 60,* 63-91.

* McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A., & Braverman, E. P. (2001). Predicting job performance using situational judgment tests: A clarification of the literature. *Journal of Applied Psychology, 86,* 730-740.

* Mercer, J. R. (1984). What is a racially and culturally nondiscriminatory test? In: *Perpectives on bias in mental testing,* ed. Reynolds, C. R. & Brown R. T. Plenum.

* Montie, J. E., & Fagan, J. F. (1988). Racial differences in IQ: Item analysis of the Stanford-Binet at 3 years. *Intelligence, 12,* 315-332.

* Naglieri, J. A., & Jensen, A. R. (1987). Comparison of Black-White differences on the WISC-R and the K-ABC: Spearman's hypothesis. *Intelligence, 11, 21-43.*

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

* Nichols, P. L. (1972). *The effects of heredity and environment on intelligence test performance in 4 and 7 year old White and Negro sibling pairs*. Doctoral dissertation, University of Minnesota.

* Nyborg, H., & Jensen, A. R. (2000). Black-white differences on various psychometric tests: Spearman's hypothesis tested on American armed services veterans. *Intelligence, 28,* 593-599.

* Office of the Assistant Secretary of Defense (Manpower, Reserve Affairs, and Logistics) (1982) Profile of American Youth: 1980 nationwide administration of the Armed Services Vocational Aptitude Battery. Department of Defense.

Pennock-Román, M. (1992). Interpreting test performance in selective admissions for Hispanic students. In K. F. Geisinger (Ed.), *Psychological testing of Hispanics* (pp. 99–135). Washington, DC: American Psychological Association.

Peoples, C. E., Fagan, J. F., & Drotar, D. (1995). The influence of race on 3-year-old children’s performance on the Stanford-Binet: Fourth Edition. *Intelligence, 21,* 69-82.

Reynolds, C. R., & Gutkin, T. B. (1981). Multivariate comparison of the intellectual performance of Blacks and Whites matched on four demographic variables. *Personality and Individual Differences , 2,* 175-180.

* Reynolds C. R., Willson, V. L., & Ramsey, M. (1999). Intellectual differences among Mexican Americans, Papagos and Whites, independent of *g. Personality and Individual Differences, 27,* 1181–1187.

* Roth, P. L., Bevier, C. A., Bobko, P., Switzer, F. S., III, & Tyler, P. (2001). Ethnic group differences in cognitive ability in employment and educational settings: A meta-analysis. *Personnel Psychology, 54,* 297–330.

* Rushton, J. P. (2001). Black–White differences on the g factor in South Africa: a ‘‘Jensen Effect’’ on the Wechsler Intelligence Scale for Children—Revised. *Personality and Individual Differences*, *31*, 1227–1232.

* Rushton, J. P. (2002). Jensen Effects and African/Colored/Indian/White differences on Raven’s Standard Progressive Matrices in South Africa. *Personality and Individual Differences, 33*, 65–70.

* Rushton, J. P., Čvorović, J., & Bons, T. A. (2007). General mental ability in South Asians: Data from three Roma (Gypsy) communities in Serbia. *Intelligence, 35,* 1-12.

* Rushton, J. P., & Jernsen, A. R. (2003). African-White IQ differences from Zimbabwe on the Wechsler Intelligence Scale for Children-Revised are mainly on the *g* factor. *Personality and Individual Differences, 34,* 177-183.

* Rushton, J. P., & Skuy, M. (2000). Performance on Raven's Matrices by African and White university students in South Africa. *Intelligence, 28,* 251−265.

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

* Rushton, J. P., Skuy, M., & Fridjohn, P. (2003). Performance on Raven's Advanced Progressive Matrices by African engineering students. *Intelligence, 31,* 123−137.

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

Schmidt, F. L., & Hunter, J. E. (1999). Theory testing and measurement error. *Intelligence, 27*, 183-198.

Schmidt, F. L., & Le, H. (2004). Software for the Hunter-Schmidt meta-analysis methods. University of Iowa, Department of Management & Organization, Iowa City, Iowa, 42242.

Spearman, C. (1923). *The nature of 'intelligence' and the principles of cognition***.** London: Macmillan.

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

* te Nijenhuis, J., Evers, A., & Mur, J. P. (2000). The validity of the Differential Aptitude Test for the assessment of immigrant children. *Educational Psychology, 20*, 99–115.

te Nijenhuis, J., & van der Flier, H. *Is the Flynn effect on g?: A meta-analysis*. Manuscript submitted for publication.

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

te Nijenhuis, J., & van der Flier, H. (1999). Bias research in the Netherlands. Review and implications. *European Journal of Psychological Assessment*, 15(2), 165–175.

te Nijenhuis, J., & van der Flier, H. (2003). Immigrant-majority group differences in cognitive performance: Jensen effects, cultural effects, or both? *Intelligence, 31,* 443-459.

* te Nijenhuis, J., & van der Flier, H. (2005). Immigrant-majority group differences on work-related measures: the case for cognitive complexity. Personality and Individual Differences, 38, 1213–1221.

te Nijenhuis, J., & Franssen, D. B. (2010). *What is the significance of test-score differences? Five psychometric meta-analyses on g loadings and IQ scores: The relation of inbreeding, visual impairment, schizophrenia, epilepsy, and giftedness with general intelligence*. Unpublished manuscript, University of Amsterdam, Amsterdam, NL.

te Nijenhuis, J., & Jongeneel-Grimen, B. (2007). *Can people get smarter: Three psychometric meta analyses and three exploratory meta-analyses on g loadings and IQ scores*. Unpublished manuscript, University of Amsterdam, Amsterdam, NL.

te Nijenhuis, J., de Jong, M.-J., Evers, A ,van der Flier, H. (2004). Are cognitive differences between immigrant and majority groups diminishing? *European Journal of Personality, 18*, 405-434.

te Nijenhuis, J., de Pater, I. E., van Bloois, R., & Geutjes, L.L. (2009). *Two psychometric meta-analyses and three exploratory meta-analyses on g loadings and IQ scores: The relation of giftedness, mental retardation, alcohol and cocaine abuse, and depression with general intelligence*. Unpublished manuscript, University of Amsterdam, Amsterdam, NL.

* te Nijenhuis, J., Tolboom, E., Resing, W., & Bleichrodt, N. (2004). Does cultural background in fluence the intellectual performance of children from immigrant groups?: validity of the RAKIT intelligence test for immigrant children. *European Journal of Psychological Assessment*, *20*, 10-26.

* U.S. Department of Labor, Manpower Administration (1970). *Manual for the USES General Aptitude Test Battery.* U.S. Employment Service.

* Valencia, R. R., & Rankin, R. J. (1986). Factor analysis of the K-ABC for groups of Anglo and Mexican American children. *Journal of Educational Measurement, 23,* 209-219.

Vernon, P. A. & Jensen, A. R. (1984). Individual and Group differences in intelligence and speed of information processing. *Personality and Individual Differences, 5,* 411-423.

* Whetzel, D. L., McDaniel, M. A., & Nguyen, N. T. (2008). Subgroup differences in situational judgment test performance: A meta-analysis. *Human Performace, 21,* 291-309.

Wicherts, J.M. (2007). *Group difference in intelligence test performance.* Unpublished doctoral dissertation, University of Amserdam, The Netherlands.