Spearman’s hypothesis tested with Raven’s
Progressive Matrices: A psychometric meta-analysis

Jasper Repko

Master thesis
University of Amsterdam
Work- and Organizational Psychology
Studentnumber: 0580015
Supervisor: Dr. Jan te Nijenhuis
May 2011

Acknowledgement
First of all I would like to thank the undergraduate students before me who worked on the research program of the combination of the method of correlated vectors and psychometric meta-analysis. Further thanks goes to dr. John Raven who was friendly and helpful to give us his views on our research. I would also like to thank dr. Arne Evers, dr. Phil Rushton, dr. Trudy Bons, dr. Ute Klehe, and dr. Henk van der Flier for statistical advice and willingness to give thought to our questions. Lastly, I thank dr. Jüri Allik for making his dataset available for our meta-analysis.
Abstract
General intelligence is the key predictor of job- and educational performance. Charles Spearman proposed in 1927 that group differences in IQ scores are a function of the cognitive complexity of these IQ scores. We conducted a psychometric meta-analysis on data from both Raven’s Standard Progressive Matrices, and Raven’s Advanced Progressive Matrices. We compared Whites, Indians, Roma, and Blacks with each other in order to test the hypothesis.

To perform the psychometric meta-analysis we used the software developed by Schmidt and Le (2004). The software corrects for four statistical artifacts namely: (1) sampling error, (2) reliability of the vector of g loadings, (3) reliability of the vector of a specific variable of theoretical interest, and (4) restriction of range of g loadings. After this correction we further corrected for (5) deviation from perfect construct validity.

Previous research found a meta-analytical correlation between g loadings and standardized group differences for test batteries of .91. In the present study, using much less studies and much smaller studies, and using item scores, we find a meta-analytical correlation of .83. These values are highly similar and suggest the meta-analytical methodology is basically sound.

The outcomes of our meta-analyses strongly support Spearman’s hypothesis. As a previous meta-analysis has already shown that g loadings and heritabilities are virtually interchangeable, the high correlation between g loadings and group differences in the present meta-analysis is in line with the hypothesis of a substantial genetic component in group differences.
General intelligence differences are associated with important life outcomes, including school achievement (Johnson, McGue, & Iacono, 2006). In a study involving tens of thousands of children, general intelligence at age eleven had a correlation of over 0.8 with scores on national tests of educational achievement five years later (Deary, Strand, Smith, & Fernandes, 2007). General intelligence is also strongly predictive of occupational attainment, social mobility, (Strenze, 2007) and job performance (Gottfredson, 1997). Moreover, people with higher general intelligence in childhood or early adulthood have better health in middle and later life, and are less likely to die young (Batty, Deary, & Gottfredson, 2007).

Much research has already been carried out on group differences in IQ scores between Blacks and Whites. Most of the research has been carried out in the United States, with the general finding being that Blacks score approximately 1.2 standard deviation lower than Whites, equivalent to about 18 points on the IQ index (Jensen, 1998). The IQ scores of North-East Asians, Indians, and Hispanics have also been well researched. There is consensus that North-East Asians score the highest on IQ tests, next Whites, then Indians, and Blacks the lowest. It is important to note that we are dealing with group averages; since there is a large overlap in distribution of IQ scores between the groups many individuals of the lowest scoring group will score higher than the majority of individuals of the highest scoring group.

With general intelligence having such impact on people’s lives, it is important to understand how and why people differ in general intelligence. Research has convincingly shown that within groups intelligence is strongly genetically determined (Jensen, 1998). However, how strong is the genetic component of the differences in general intelligence between groups? The position that group differences are one hundred percent caused by the environment does not seem to be supported (Hunt, 2010), but there is no consensus as to the size of the genetic component. A meta-analysis of Spearman’s hypothesis – group differences on IQ tests are a function of the cognitive complexity of these IQ tests – acts as a test of the size of the genetic component of group differences, as we will describe below.

General Intelligence (g)

As stated by te Nijenhuis and Dragt (2010), 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) (te Nijenhuis, de Pater, van Bloois, & Geutjes, 2009).

Hierarchical Intelligence Model

Te Nijenhuis and Dragt (2010) continue by stating that 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) (te Nijenhuis & Dragt, 2010).

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) (te Nijenhuis & Dragt, 2010).
Spearman’s Hypothesis

Charles Spearman (1927) hypothesized that the relative size of the mean Black/White difference (in standardized scores) on various mental tests is a direct function of the tests’ different loadings on the general factor, psychometric $g$, the highest order common factor in all complex tests of cognitive ability. Since then this hypothesis has not only been extensively tested, but the hypothesis has also been expanded to group differences in general. However, the Black/White difference remains the most-studied topic and Jensen (1998) concludes that the Black/White difference mainly consists of a difference in $g$. Research has shown that differences between various other groups can be contributed to differences in $g$ as well.

Recently, a psychometric meta-analysis of IQ batteries by te Nijenhuis and Dragt (2010) showed a strong relation between group differences and complexity: a true correlation 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. Their 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.

Te Nijenhuis and Dragt (2010) describe how Spearman’s hypothesis has also been studied using other methods than intelligence batteries. 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 clear 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) (te Nijenhuis & Dragt, 2010).

Second, te Nijenhuis and Dragt (2010) describe how Spearman's hypothesis has also been studied using Situational Judgment Tests (SJTs) 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,
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 group 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 competencies. 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 group differences in cognitive ability in employment and educational settings show differences that are largest on the most g-loaded measures (te Nijenhuis & Dragt, 2010).

Thirdly, Spearman’s hypothesis has also been tested using the item scores of the Raven’s Standard Progressive Matrices (SPM) and the Raven’s Advanced Progressive Matrices (APM). The Progressive Matrices Test is one of the most widely used measures of cognitive ability (Raven, Court, & Raven, 1996). Rushton and colleagues conducted six studies on Spearman’s hypothesis tested with the Raven Progressive Matrices. The first study was carried out at the University of the Witwatersrand and the Rand Afrikaans University in Johannesburg in South Africa (Rushton & Skuy, 2000). They tested 173 African and 136 White students on the SPM. The second study that he conducted was again at the University of the Witwatersrand
where they tested 198 African students, 86 White students, and 58 Indian students on the SPM (Rushton, Skuy, & Fridjhon, 2002). With the same colleagues he conducted the third study, again at the University of the Witwatersrand (Rushton, Skuy, & Fridjhon, 2003). Here he gave the APM to 187 African students, 67 White students, and 40 Indian students. In 2004 he carried out another test using the APM at the University of the Witwatersrand (Rushton, Skuy, & Bons, 2004) and this time he tested 177 African students, 72 White students, and 57 Indian students. In addition to these five studies Rushton (2002) also re-examined the data of the study by Owen (1992) who gave the SPM to 1093 African, 1056 White, 778 Colored, and 1063 Indian 14-year-olds. The sixth study that Rushton conducted was among Serbian Roma (Rushton, Čvorović, & Bons, 2007). They compared the data of 231 Roma with the data of a sample of African students reported by Owen (1992). In all the six studies carried out by Rushton substantial mean score differences between the groups are found that confirm Spearman’s hypothesis.

Fourth, beside IQ batteries, reaction time measures, and Raven's Progressive Matrices, te Nijenhuis and Dragt (2010) describe how 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. The exploratory meta-analysis of te Nijenhuis and Dragt (2010) shows that language-biased subtests underestimate IQ for minority groups, but virtually all effects are small.

In sum, te Nijenhuis and Dragt (2010) conclude that Spearman’s hypothesis was confirmed in the large majority of comparisons of various groups and for all assessment instruments studied.

*g loadings and heritabilities*

Causes of group differences in intelligence were and are still being discussed intensively (Brody, 1992; Galton, 1869; Jensen, 1998; Loehlin, Lindzey, & Spuhler, 1975; Nisbett, 2010; Rushton & Jensen, 2005ab, 2010; Vernon, 1979). Rushton and Jensen (2005, 2010) state that differences in intelligence between Blacks and Whites in the US are at least 50% heritable, based on several lines of research. First, in every part of the world, Whites have higher IQ scores than Blacks. Second, B/W differences
in IQ appear early: Black and White three-year-old children already differ fifteen IQ points. Third, in most studies heritability of intelligence is equal in groups of Blacks and Whites. Fourth, research has clearly shown a substantial positive correlation between brain volume and intelligence and group differences in brain volume are large. Fifth, studies of, respectively, Black and Korean children adopted into White homes show that their IQs do not resemble those of their adoptive parents, but those of their biological parents. So, the adopted Black children had lower IQs than their adoptive parents, and adopted Korean children had higher IQs than their adoptive parents. Sixth, groups of mixed-race individuals have mean scores intermediate to unmixed groups of Blacks and Whites. Seventh, children of two high-IQ parents show lower IQ scores than their parents – an effect known as regression to the mean. The regression effects are much stronger for Black families than for White families, implying there is regression towards different means for groups with different mean IQs.

Brody (1992) is not convinced there is a genetic component to Black/White differences. First, there are only few studies on the heritability of intelligence in Black samples, so it is not possible to draw strong conclusions regarding equal heritability for Blacks and Whites. Second, there may be environmental influences that Blacks commonly encounter, but that Whites are not likely to encounter. Third, all the arguments in favor of the genetic hypothesis of B/W differences are indirect; direct tests of the hypothesis are impossible, yes even unethical. Nisbett (2010) even claims that environmental factors fully explain B/W differences and that the genetic contribution is nil.

Notwithstanding the large differences of opinion, there appears to be some kind of consensus on the topic of the causes of group differences. Snyderman and Rothman (1987) surveyed over one thousand experts in behavioral genetics and psychometrics and they found that the majority believed the B/W IQ gap to be a product of both genetic and environmental causes. So, Nisbett’s position that the genetic contribution to group differences is nil appears to be quite extreme among experts.

Jensen (1987) shows that g loadings of tests correlate highly with their heritabilities in two datasets using the WAIS. Spitz (1988) reports a) g loadings of subtests, b) scores of mentally retarded, and c) heritability coefficients, and the correlations between a and b and between b and c, but does not actually compute the
correlation between $g$ and $h^2$. However, as the correlations between a and b and between b and c are high it seems logical that a and c will also correlate positively (Van Bloois, Geutjes, te Nijenhuis, de Pater, & Grimen, 2009). Pedersen, Plomin, Nesselroade, and McClearn (1992) report a correlation of .77 between $h^2$ and $g$, whereas Rijsdijk, Vernon, and Boomsma (2002) report a value of .43 (Van Bloois et al., 2009). Te Nijenhuis and Jongeneel-Grimen (2007) show there is a meta-analytical correlation between $g$ loadings and heritabilities of +1, based on a large total $N$.

Jensen (1998) has shown that a test’s $g$ loading is the best predictor of inbreeding depression scores, brain evoked potentials, brain pH levels, brain glucose metabolism, as well as nerve conduction velocity, reaction time, and other physiological factors. These correlations are in line with the heritable and biological reality of $g$. The studies cited above convincingly show that group differences in IQ are more pronounced on highly $g$-loaded tests than on tests with a low $g$ loading.

So, the rank order of IQ subtests’ cognitive complexity, heritability, and group differences appears to correlate highly. This strongly suggests that the main source of the group differences across various cognitive tests is essentially the same as the main source of differences between individuals within each group, namely $g$. So, this implies there is a substantial genetic component in group differences. Of course, the conclusion is based on correlations and not on the outcomes from experimental studies, so a strong confirmation of Spearman’s hypothesis does not prove there is a substantial genetic component in group differences, but at best makes that conclusion plausible.

**Research Question**

Te Nijenhuis and Dragt (2010) report in their meta-analysis of IQ batteries a meta-analytical rho of .91 between true $g$ and true $d$. If we assume that the meta-analytical corrections in their study are carried out correctly, a meta-analysis of studies of Raven’s Progressive Matrices at the item level using comparable statistical corrections should lead to a comparable value of the meta-analytical rho. The research question of this study is therefore whether a meta-analysis of Spearman’s hypothesis using Raven’s data results in a value of rho that is quite close to 1. Such an outcome would be in line with a substantial genetic component in group differences.
Method

Instruments

There are four versions of the Raven’s: the Standard Progressive Matrices (SPM) for the ages of 6 years to adulthood; the Colored Progressive Matrices, an easier version of the test designed for children aged 5 through 12; the Advanced Progressive Matrices (APM), a harder version of the test designed for older adolescents and adults with higher ability; and the Standard Progressive Matrices Plus (SPM plus), an extended version of the SPM offering more discrimination among more able young adults.

The SPM consists of 60 diagrammatic puzzles, each with a missing part that the test taker attempts to identify from several options. The 60 puzzles are divided into five sets (A, B, C, D, and E) of 12 items each. To ensure sustained interest and freedom from fatigue, each problem is boldly presented, accurately drawn, and, as far as possible, pleasing to look at. No time limit is set and all testees are allowed to complete the test. As an untimed capacity test, and even as a 20-min speed or efficiency test, the SPM is usually regarded as a good measure of the nonverbal component of general intelligence rather than of culturally specific information. The total score is also a very good measure of g, the general factor of intelligence, at least within Western countries (Jensen, 1980).

The Colored Progressive Matrices (CPM) is designed for young children, elderly, and people with moderate or severe learning difficulties. The test consists of 36 items. The first 12 items are the same as the first 12 items from the SPM (set A); the following 12 items are specially designed for the CPM. The last 12 items are again items from the SPM, this time items 13-24 (set B). Most of the items are presented on a colored background in order to keep the attention of the participants. The last couple of items from the test are, however, presented in black and white. If the participants manage to complete the test without too much difficulty they can continue without any problems with sets C, D, and E of the SPM.

The APM is a more difficult version of the Raven Progressive Matrices. The advanced form of the Matrices contains 48 diagrammatic puzzles, presented as one set of 12 (set 1), and another of 36 (set 2). Set 1 serves as an introduction, and set 2 is used to test the participants and compare scores. Items are presented in black ink on a white background, and become increasingly difficult as progress is made through each
set. These items are appropriate for adults and adolescents of above average intelligence. Whenever we refer to the APM in this paper, we refer to set 2 of the APM.

The SPM Plus is the extended version of the SPM and is meant to restore the discriminative ability of the test (Raven, 1998). The SPM Plus was introduced as a revised version of the SPM. It consists of new items, most of which have been equated to match the old items in difficulty, but some are more difficult than any that appear on the SPM.

In this paper we have only included studies using the SPM and the APM. For practical reasons we decided not to include studies using the short 20-minute version of the SPM, the CPM, and the SPM Plus: we found so few datasets that we could not test Spearman’s Hypothesis combining two datasets in a meaningful manner.

Next, the classical method for testing Spearman’s hypothesis, as reported by Jensen, is described. Then we describe the psychometric meta-analytical approach, which makes extensive use of the work by te Nijenhuis and Dragt (2010).

*The classical method for testing Spearman's hypothesis*

Te Nijenhuis and Dragt (2010) cite Jensen (1993) who states that seven methodological requirements for the testing of Spearman's hypothesis have to be met when using the subtests of IQ batteries:

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 coefficient indicates a high degree of similarity, the $g$ loadings of the different groups should be averaged.
6. To rule out the possibility that the correlation between the vector of $g$ loadings ($V_g$) and the vector of mean differences between the groups, or effect sizes ($V_{ES}$), is strongly influenced by the variables' differing reliability coefficients, $V_g$ and $V_{ES}$ should be corrected for attenuation by dividing each value by the square root of its reliability.
7. The test of Spearman's hypothesis is the Pearson correlation ($r$) between $V_g$
and $V_{ES}$. To test the statistical significance of $r$, Spearman's rank order correlation ($r_s$) should be computed and tested for significance.

As stated above, these seven requirements apply to test batteries. When using Raven’s data at the item level in a psychometric meta-analysis the requirements need to be changed. With concern to Jensen’s requirement numbers 3, 4, and 5 it should be noted that we do not work with subtest scores but with item scores, and that item scores have much lower reliabilities than subtest scores. We make the choice of accepting much lower values of the congruence coefficient.

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 size becomes so large that significance testing is a waste of time (te Nijenhuis & Dragt, 2010).

Specific Criteria for Inclusion

The data that we used for our meta-analysis had to be reported or be available at the item level enabling us to compute the correlations between complexity and group differences ourselves. We used both data published at the item level and Raven data in SPSS data files with scores of groups of individuals available at the item level.

Five specific decision rules for selecting relevant studies for this meta-analysis were used. At first, only studies reporting a test of Spearman’s hypothesis using the Raven’s Progressive Matrices, computing a correlation between standardized group differences and $g$ loadings, were included in the analysis however. Second, because we knew at the outset there were only few studies explicitly testing Spearman’s hypothesis, we also included studies reporting the proportion of a sample selecting the correct answer on the items ($d$), and studies reporting correlation between the item score and the total score ($g$), and studies reporting both $d$ and $g$. Aside from the published studies, either on Spearman’s hypothesis or reporting usable data, we also used data sets provided to us by various well-known authors. To make sure that the same methods were applied to all datasets in the meta-analysis, we made use of correlations that we computed ourselves instead of the correlations reported in the original articles. Our choice of decision rules imply that many studies using the Raven’s could not be used because the data were not reported at the item level: when participants are tested on the Standard Progressive Matrices, the Colored Progressive
Matrices, or the Advanced Progressive Matrices, in the large majority of cases only the percentile scores and group averages are reported.

Third, following the manual of the APM (Raven, 1998) and Rushton, we decided to exclude studies from our meta-analysis where the timed version of the APM was taken under time constraints with the time limit set to less than 30 minutes. The outcomes of the White sample of the Vigneau and Bors (2005) study does not correlate well with the outcomes of the studies using representative White groups. However, the time limit was set to 30 minutes and therefore we included the study in our analysis. We hypothesized that this dataset would show up as extreme outlier in the meta-analysis, so it would be left out as the meta-analysis advanced.

Unfortunately we were unable to include a large sample from the University of Amsterdam (UvA) in our analysis. First-year psychology students at the UvA are obliged to take a series of tests and questionnaires. One of the tests that is taken is Raven’s APM. We were allowed access to the last ten years of test data. Given the fact that every year approximately 450 students take these tests we had access to a sample of about 4500 students. However, the time-limit was set to 22 minutes so the students were unable to try all of the 36 items which yielded incomplete data. None of our other datasets included a group where the time-limit was also set to 22 minutes, so we were unable to match the group of first-year psychology students and this rendered the data unusable.

Fourth, we decided to exclude the White German sample from the Raven Advanced Progressive Matrices manual (Court & Raven, 1982), the White sample from Vejleskov (1968), the White sample from Moran (1986), and the White sample from Forbes (1964) because we did not have comparable samples of either Blacks or Indians to match these White samples with. Furthermore we decided to exclude the data of the $g$ loadings from the German sample from all our analyses (Raven, 1982) because these $g$ loadings show weak correlations with all the other available $g$ loadings. In contrast, all the other available $g$ loadings without exception correlate highly with each other. We speculate that the $g$ loadings of the White German sample were calculated with a different technique from the one that we are using. The values of $d$ of the German sample (Court & Raven, 1982) do seem usable, but since we did not have a group to match them with we did not use this part of the data either.

Fifth, according to John Raven (personal communication, 2010) there appears to be a ceiling effect taking places in some samples taking Raven’s Standard
Progressive Matrices. Ceiling effects happen when the assignment is too easy, so most people get the answer right, resulting in many perfect scores and reduced variance. A ceiling effect makes it difficult to assess the ability of the people who took the assignment and because of that it is hard to discriminate between these people. The dataset of Lynn, Allik, and Irwing (2004) appears to show such a ceiling effect for all of their age categories ranged 12 to 18, meaning that we could not compare different age categories of this sample with each other to calculate reliabilities for both $d$ and $g$. We were however able to use the dataset to compute score differences for multiple age categories between Blacks and Whites. We found that as soon as we matched corresponding age categories with each other it provided useable results.

**Sample Size: Harmonic Means**

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. Following te Nijenhuis and Dragt (2010), we “therefore for all comparisons 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:”

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

**Corrections for Artifacts**

Following te Nijenhuis and Dragt (2010) 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 (te Nijenhuis & Dragt, 2010).
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 (te Nijenhuis & Dragt, 2010).

Correction for Reliability of the Vector of g Loadings

As mentioned by te Nijenhuis & Dragt (2010), the values of $r(d \times g)$ are attenuated by the reliability of the vector of g loadings. 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 12 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 17 to 29 years.

First, we made nine pairings based on background variables (see Figure 1). Next, we decided to exclude the pairings with sample sizes that deviated more than one hundred percent from each other. This resulted in the exclusion of one sample, so there were only eight combinations left (see Figure 2): five of them taking the SPM, and three taking the APM. This is a very limited number of data points to estimate the distribution of reliabilities for the SPM and the APM separately. In order to obtain a higher quality estimate of the distribution of reliabilities we decided to combine the data points from both tests in the one graph which is presented on the following page. Presenting the SPM and the APM separately would not have painted such a clear picture as they do combined.

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 $r(g \times g)$ 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 (see Figure 1 and Figure 2).
Figure 1

*Distributions of Reliability of g, all Samples Included*
Figure 2
Distributions of Reliability g After Exclusion of Data point Consisting of Pairings of Datasets of Sample Sizes That Deviate More Than one 100 Percent From Each Other
Correction for Reliability of the Vector of the Second Variable

The values of $r (g \times d)$ 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 (Te Nijenhuis & Dragt, 2010).

For both the comparisons using children and the comparisons using adults we chose samples that were highly comparable with regard to age. We increased the number of data points by making combinations of data from different studies when two data points based on combinations of data were compared. Using this technique we made sure there was no overlap between the datasets.

Correction for Restriction of Range of $g$ Loadings

Te Nijenhuis and Dragt (2010) indicate that the values of $r (g \times d)$ can be attenuated by the restriction of range of $g$ loadings of subtests, and this principle also applies to items. 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 $u = SD_{\text{study}}/SD_{\text{ref}}$ (Te Nijenhuis & Dragt, 2010).

The average standard deviation found in this study for the Standard Progressive Matrices is 0.120, and for the Advanced Progressive Matrices the average standard deviation is 0.114. Because the average standard deviations of Raven’s SPM and APM are highly similar, we decided not to correct for the restriction of range of $g$ loadings.

Correction for Deviation from Perfect Construct Validity

Te Nijenhuis and Dragt (2010) state that the deviation from perfect construct validity in $g$ attenuates the values of $r (g \times d)$. 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 $r \ (g \times d)$ are attenuated by psychometric sampling error in each of the batteries from which a $g$ factor has been extracted (te Nijenhuis & Dragt, 2010).

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 $r \ (g \times d)$ after correction for the first four artifacts (te Nijenhuis & Dragt, 2010).

Previous studies (te Nijenhuis & Franssen, 2010; te Nijenhuis, & Jongeneel-Grimen, 2007; te Nijenhuis, de Pater, van Bloois, & Geutjes, 2009; te Nijenhuis & van der Flier, submitted; te Nijenhuis, van Vianen, & van der Flier, 2007) used a conservative value of .90 as a basis to limit the risk of overcorrection. However, te Nijenhuis and Dragt (2010) computed a $g$ score based on 24 subtests and various $g$ scores based on combinations of eleven subtests (the most common number of subtests in a battery); this yielded an average correlation between their estimate of “true $g$” and the other $g$’s of 0.925. The new method for computing the correction for imperfectly measuring $g$ from te Nijenhuis and Dragt (2010) shows that the correction used before was too strong. In all previous studies a correction of ten percent was applied to compute rho-5, but in this paper te Nijenhuis and Dragt (2010) rounded of
the value for the correction of the fifth artifact to 7.5% for the computation of rho-5. However, te Nijenhuis and Dragt (2010) worked on test batteries, whereas in the present meta-analysis the focus is on the Raven’s. Based on our reading of the literature – for instance the dataset from the Minnesota Twin Study by Bouchard (Bouchard, Lykken, McGue, Segal, & Tellegen, 1990) – we estimate that the total score on the Raven’s has a g loading of about .75. We remind the reader that the correlation of the item score with the Raven’s total score is an estimate of an item’s g loading. So this suggest that for research at the item level a correction of 7.5% has to be combined with a correction of 25%, yielding a total correction of 32.5%.

Searching and Screening Studies

To identify studies for inclusion in the meta-analysis, both electronic and manual searches were conducted for studies in which Spearman's hypothesis was tested that contained data on Raven’s Progressive Matrices at the item level. Four 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, PubMed, and google.scholar was conducted. The following combinations were used to conduct the searches: any keyword that contains the words Raven Progressive Matrices, Standard Progressive Matrices, Advanced Progressive Matrices, Spearman’s hypothesis, item-total correlation, and '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, we also checked articles by the following authors: John Raven, Phil Rushton, and Richard Lynn. Third, reference lists of all currently included empirical studies were checked to identify any potential articles that may have been missed by earlier search methods. Fourth, all the articles mentioned in the reference lists of the meta-analyses by Lynn and Irwing (2004), Brouwers, van de Vijver, and van Hemert (2008), and Wicherts, Dolan, Carlson, and van der Maas (2010) were searched for and, when found, checked for relevance for our meta-analyses. Fourth, the dissertation by Zaaiman (1998) was studied for data. Fifth, several well-known researchers who have conducted research on Spearman's hypothesis were contacted in order to obtain any additional articles or supplementary information.

Estimating g

The total score on the Raven’s is a good measure of g, so the item-total
correlation gives an estimate of the $g$ loadings of the items on the test (Jensen & Weng, 1994). In this paper we used both the $g$ loadings provided by the authors in the various papers, and the $g$ loadings that we computed ourselves using data supplements. It was impossible to compute the $g$ scores of every available study because the necessary data were not reported in many of the articles or the data supplements used in our study, so in many cases we had to use the $g$ scores that were calculated by the authors of the articles.

Correlations between various mental tests range from slightly greater than 0 to slightly less than 1, but they are always positive except for sampling error or statistical artifacts (Jensen, 1987a). Until someone can devise a cognitive test that has a true negative correlation with other mental tests, which no one has yet succeeded in doing, it can be accepted as a fact that all types of mental tests are positively correlated (Jensen, 1992). So, when negative $g$ loadings of items were reported in the articles that have been used for this meta-analysis, we did not exclude these items, but we changed the value of the $g$ loadings to .00.

We used in the large majority of the cases the $g$ loadings of the White group. Exceptions were made for the following cases: the $g$ of the Roma was used to compute the correlation between the Roma sample of Rushton, Čvorović, and Bons (2007) and the Black sample of Rushton and Skuy (2000). For the correlation between the Whites of Lynn, Allik, and Irving (2004) and the Blacks of Rushton and Skuy (2000) we used the $g$ of the White sample from Rushton and Skuy (2000). For the correlation between the Whites of Lynn, Allik, and Irving (2004) and the Blacks of Rushton, Skuy, and Fridjhon (2002) we used the $g$ of the White sample from Rushton, Skuy, and Fridjhon (2002). The $g$ of the Whites from Rushton (2002) was used to compute the correlation between the Whites of Lynn, Allik, and Irving (2004) and the Blacks of Rushton (2002).

A general aim of the meta-analysis was to get the best estimates of $d$ and $g$ as possible. A general principle of psychometric meta-analysis is that combining a substantial number of datasets reduces error and increases the quality of the estimates. We therefore computed an aggregated $g$ for, respectively, the Standard Progressive Matrices and the Advanced Progressive Matrices by combining all the available data of these tests. However, the $g$ loadings of teenagers and the $g$ loadings of adults yielded a low correlation, so we concluded that they were not comparable. For the SPM we therefore decided to compute both an aggregated $g$ loading for groups with a
mean age of 14, and one for groups with the lower bound of the age range being 18 years. For the APM the only data that were available to us were of adults, so we only computed one aggregated \( g \) loading for this measurement instrument. The three different aggregated \( g \)s based on several datasets were used to compute the correlation between \( d \) and \( g \), in addition to the \( g \)s that were based upon individual datasets. So for every \( d \) we calculated two correlations, one with the \( g \) of the White group (unless as stated we used a different \( g \)), and one with the matched aggregated \( g \).

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. The only exception was the \( g \) of the Roma taking the SPM (Rushton, Čvorović, & Bons, 2007), a group with a minimum age of 17. This \( g \) was found to be incomparable to other samples with an age of above 17, but highly comparable to the \( g \) of the group with an average age of 14 years old. Therefore we decided to add the \( g \) of the Roma to the SPM group with an average age of 14 years old. Aggregated \( g \) loadings were obtained by combining all the available \( g \) loadings to form one aggregated \( g \) for every age group. This was done by multiplying the \( g \) loadings of every group by the total participants of that group, adding the multiplied scores of the groups up and dividing that number by the sum-total of the participants used to calculate the specific aggregated \( g \). This way, the largest datasets were weighted most strongly. This was done for every single item.

**Estimating \( d \)**

Following Rushton (Rushton & Skuy, 2000; Rushton, Skuy, & Fridjhon 2002, 2003; Rushton, Skuy, & Bons, 2004) first of all the pass-rates from every group were changed to standard scores by use of a Z-transformation before being subtracted from each other. It should be noted that the descriptions in Rushton’s Method section do not always match the outcomes. However, our computations are in line with Rushton’s outcomes. Score differences between two groups (\( d \)) were computed by subtracting the mean proportion correct of the lower scoring group from the mean proportion correct of the higher scoring group (to generally obtain positive scores). Groups were matched based on average age and education.

**Reliability of \( d \)**

To estimate the reliability of the \( d \) vector we compared the \( ds \) of two comparable samples, for all possible combinations. Moreover, a few of the samples were combined to create groups with larger sample sizes resulting in estimates of
reliability for a large $N$. The analyses were carried out in a stepwise fashion. First of all we made combinations based on solely the age category. Second, we excluded the combinations where the value of the harmonic $N$ of the two samples deviated more than a hundred percent from each other. Third, we excluded all the negative correlations to smoothen the data (see Figure 3).
Figure 3

Distribution of Reliability of $d$
Choice of Correlation Between Item and Total Score, and Between $d$ and $g$

First, we will address the choice of the correlation coefficient for computing the item-total correlation. Second, we will argue our choice of correlation coefficient when computing the correlation between the $g$ vector and the $d$ vector.

When one of two variables is continuous, and the other is dichotomous, the biserial correlation coefficient should be used. If no underlying continuum is assumed, or if the assumption of normality is questionable, the point-biserial coefficient of correlation should be used (Guion, 1998, p. 327). The point-biserial correlation coefficient will be used to compute the correlation between the item score and the total score ($= g$ loading), because the scores on the item are dichotomous, and the total scores are continuous. Moreover, the use of the point-biserial correlation was also advised to us by dr. Arne Evers (personal communication, June 1st 2010). In the SPSS program we used the values of the Pearson correlation coefficient because Guilford and Fruchter (1978, p. 308-309) state that the value of Pearsons correlation is equivalent to the value of a point-biserial $r$.

With concern to $r(d \times g)$, since both our $d$ values and our $g$ values are continuous, we have to use either Spearman’s $\rho$, or Pearson’s $r$. Spearman’s $\rho$ is a non-parametric measure of statistical dependence between two variables. It assesses how well the relationship between two variables can be described using a monotonic function. The Spearman correlation coefficient is simply the Pearson correlation coefficient between the ranked variables.

Pearson’s $r$ is used when both variables have interval scales of measurement. The three assumptions underlying the use of Pearson’s $r$ are the following:

1) The relationship between $X$ and $Y$ should be essentially monotonic and preferably linear. Linear correlations involve analysis of straight-line relationships between two variables. Linear correlations can range between -1 (perfectly negative) and +1 (perfectly positive), with 0 indicating no straight-line relationship. This means in practice that $r$ describes a relationship poorly when the trend line increases and then decreases, or vice versa (is nonmonotonic) (Nunnally & Bernstein, 1994, p. 129).

2) The relationship must be homoscedastic so that the spread (errors of estimate) about the best-fitting straight line is approximately the same at all levels of $X$ and $Y$, rather than heteroscedastic, where the spread is much greater at certain levels than others (Nunnally et al., 1994, p. 129-130). The assumption of
homoscedasticity simplifies mathematical and computational treatment. Serious violations in homoscedasticity result in overestimating the goodness of fit as measured by the Pearson coefficient, which means that modest violations of homoscedasticity pose no serious problem.

3) Error affecting each of the variables must be normally distributed (not necessarily the variables themselves) if inferential test are to be used. Even though this assumption is not necessary to simply describe the relation, extreme skewness can lead to other misleading results even in describing the relationship (Nunnally et al., 1994, p. 130). So, results will only be compromised in case of extreme skewness.

Nunally and Bernstein (1994) state that there is nothing to prevent the use of $r$ even if the shapes of the distributions are markedly different, the relationship is far from linear, and/or the spread varies along the line. Unless these assumptions are seriously violated, no real problem in interpretation arises (Nunnally et al., 1994, p. 130).

Little is known about the distribution of $g$s and $d$s using item scores. To test if there are marked differences between the outcomes for the two correlation coefficients we decided to calculate both Spearman’s $\rho$ and Pearson’s $r$ for every relationship studied.

We checked for linearity and in the large majority of cases clear linear relationships emerged; the only clear non-linear relationships were found in two of the smallest datasets and it’s well-known that outcomes from small samples are much more unreliable than outcomes from large samples. We also checked for homoscedasticity and found that in the large majority of cases the spread around the regression line were not dissimilar for different score ranges. To check if the error affecting each of the variables is normally distributed we used the Shapiro-Wilk test. Most of the data were found to be normally distributed with a minority of cases not far from a normal distribution. We found no evidence of extreme skewness among our samples. So, these findings strongly support the use of Pearson correlations.

**Results**

We conducted a series of seven meta-analyses with the data that we gathered. The first four meta-analyses were conducted using the original $g$ loadings from individual
datasets. The last three meta-analyses were conducted using the aggregated $g$ loadings matching the age range of the comparisons of $d$.

Because the SPM and APM are very similar tests and the value of the $SD$ of the $g$ loadings are nearly the same – .120 for the SPM and .114 for the APM – we combined the data points from the two tests. For our analyses we used the meta-analytical program from Schmidt and Le (2004) to calculate the meta-analytical correlations between $d$ and $g$, based upon the corrections for the five statistical artifacts mentioned in Method.

For the first analysis we included all the data that we gathered. Figure 4.1 shows all the 34 combinations of $r(dxg)$ that were used in this meta-analysis.

![Figure 4.1](image)

*Figure 4.1*

*34 Combinations of $r(dxg)$ of the SPM and APM*
Table 1

Meta-analytical output of 34 Combinations of r(dxg) of the SPM and APM

<table>
<thead>
<tr>
<th>$K$</th>
<th>$N_H$</th>
<th>$r$</th>
<th>$SD_r$</th>
<th>rho-4</th>
<th>$SD_{\text{rho-4}}$</th>
<th>rho-5</th>
<th>%VE</th>
<th>80% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>34</td>
<td>7662</td>
<td>.47</td>
<td>.31</td>
<td>.96</td>
<td>.57</td>
<td>1.27</td>
<td>12%</td>
<td>.22 - 1.69</td>
</tr>
</tbody>
</table>

The outcomes of the first analysis are reported in Table 1. $K$ stands for the number of studies, $N_H$ stands for the total harmonic $N_H$, $r$ stands for the bare-bones correlation between $d$ and $g$, $SD_r$ is the standard deviation of this correlation. Rho-4 stands for the correlation between $d$ and $g$ corrected for the first four statistical artifacts, and $SD_{\text{rho-4}}$ is the standard deviation of this correlation. Rho-5 stands for the correlation between $d$ and $g$ corrected for all five of the statistical artifacts with the %VE showing the percentage explained variance between the number of studies ($K$). The 80% CI stands for the credibility interval. The correlation between $dxg$ is 1.27 and the percentage explained variance is only 12.

The very low percentage of variance explained suggest that there are strong moderators. The first meta-analysis consisted of data points of $r(dxg)$ in the age range 14-30. Since we had just four data points with an average age of 14, and 30 data points with an age range of 17-30 we decided to exclude these four data points (see Table 8) from our second meta-analysis (Figure 4.2). In other words, we tested age as a moderator.
The meta-analytical outcomes in Table 2 show a correlation between $d_{xg}$ of .48 and a percentage explained variance of 54. The increase in percentage variance explained clearly shows that age acted as a moderator.

As the next step, we decided to exclude the data points using the study of Vigneau and Bors (2005) (see Table 8) for $d_{xg}$, $g_{xg}$, and $d_{xd}$ because the sample size of the White group from Vigneau and Bors (2005) deviated too much from the Black samples which they were compared with. The reason we didn’t exclude these data points in the first place is because of the general principle from a Schmidt and Hunter-style psychometric meta-analysis to include all data, and to exclude extreme outliers.
and outliers at a later stage. Excluding the data points using the Vigneau and Bors (2005) study left 28 data points for the third meta-analysis, and these data points are shown in Figure 4.3.

Figure 4.3
28 Combinations of $r(dxg)$ of the SPM and APM

Table 3
Meta-analytical output of 28 Combinations of $r(dxg)$ of the SPM and APM

<table>
<thead>
<tr>
<th>K</th>
<th>$N_H$</th>
<th>$r$</th>
<th>$SD_r$</th>
<th>rho-4</th>
<th>$SD_{rho-4}$</th>
<th>rho-5</th>
<th>%VE</th>
<th>80% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>28</td>
<td>2830</td>
<td>.18</td>
<td>.12</td>
<td>.44</td>
<td>.13</td>
<td>.58</td>
<td>81%</td>
<td>.28 -.61</td>
</tr>
</tbody>
</table>

The meta-analytical outcomes in Table 3 show a correlation between $dxg$ of .58 and a percentage explained variance of 81.

For the fourth and last meta-analysis using the original $g$ loadings we excluded two further outliers (see Table 8). The data points of the Whites and Indians, and Indians and Blacks from Rushton, Skuy, and Bons (2004) correlated negatively with
the g loadings of the White sample from Rushton et al. (2004). Since both the negative correlations of $dxg$ came from the same sample we decided to exclude that sample from our fourth meta-analysis. This left a total of 26 combinations of $dxg$, as displayed in Figure 4.4.

![Figure 4.4](image)

**Figure 4.4**

26 Combinations of $r(dxg)$ of the SPM and APM

Table 4

<table>
<thead>
<tr>
<th>Meta-analytical output of 26 Combinations of $r(dxg)$ of the SPM and APM</th>
</tr>
</thead>
<tbody>
<tr>
<td>K</td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>26</td>
</tr>
</tbody>
</table>

The meta-analytical outcomes in Table 4 show a correlation between $dxg$ of .62 and a percentage explained variance of 104. Leaving out the outliers strongly increased the amount of variance explained.

The next series of three meta-analyses was conducted using aggregated $g$
loadings. We did not correct for the second statistical artifact, the reliability of the vector of $g$ loadings, because the aggregated $g$ that we use for both the SPM as the APM is based on such a large sample that is has virtually perfect reliability.

For the first of these three meta-analyses we included all the combinations of $d \times$ aggregated $g$. This resulted in the 34 data points shown in Figure 5.1.

![Figure 5.1](image)

**Figure 5.1**

34 Combinations of $r(d \times$ aggregated $g)$ of the SPM and APM

**Table 5**

*Meta-analytical output of 34 Combinations of $r(d \times$ aggregated $g)$ of the SPM and APM*

<table>
<thead>
<tr>
<th>$K$</th>
<th>$N_H$</th>
<th>$r$</th>
<th>$SD_r$</th>
<th>rho-4</th>
<th>$SD_{rho-4}$</th>
<th>rho-5</th>
<th>%VE</th>
<th>80% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>34</td>
<td>7662</td>
<td>.59</td>
<td>.32</td>
<td>.70</td>
<td>.35</td>
<td>.93</td>
<td>12%</td>
<td>.25 - 1.15</td>
</tr>
</tbody>
</table>

The meta-analytical outcomes in Table 5 show a correlation between $d \times$ aggregated $g$ of .93 and a percentage explained variance of 12.

Since the percentage explained variance of the 34 data points was quite low,
we decided to test for moderators. For the second of this series of meta-analyses we excluded the samples where the age of the participants was lower than 17. This excluded four data points of $d \times \text{aggregated } g$ (see Table 8), leaving the 30 data points displayed in Figure 5.2.

![Figure 5.2](image)

**Figure 5.2**

30 Combinations of $r(d \times \text{aggregated } g)$ of the SPM and APM

Table 6

<table>
<thead>
<tr>
<th>$K$</th>
<th>$N_{ij}$</th>
<th>$r$</th>
<th>$SD_r$</th>
<th>rho-4</th>
<th>$SD_{\text{rho-4}}$</th>
<th>rho-5</th>
<th>%VE</th>
<th>80% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>3365</td>
<td>.34</td>
<td>.30</td>
<td>.47</td>
<td>.38</td>
<td>.62</td>
<td>13%</td>
<td>-.02</td>
</tr>
</tbody>
</table>

The meta-analytical outcomes in Table 6 show a correlation between $d \times \text{aggregated } g$ of .62 and a percentage explained variance of 13.

The percentage explained variance for these 30 data points is still very low. Inspection of the scatter plot shows there are outliers that cause the low amount of
explained variance. The two data points based upon data from the study from Vigneau and Bors (2005) were identified as the extreme outliers. For the third and final meta-analysis of this series we excluded these outliers (see Table 8). This resulted in a total of 28 combinations of $d \times \text{aggregated } g$ displayed in Figure 5.3.

![Figure 5.3](image)

**Figure 5.3**

28 Combinations of $r(d \times \text{aggregated } g)$ of the SPM and APM

**Table 7**

Meta-analytical Output of 28 Combinations of $r(d \times \text{aggregated } g)$ of the SPM and APM

<table>
<thead>
<tr>
<th>$K$</th>
<th>$N_{ij}$</th>
<th>$r$</th>
<th>$SD_r$</th>
<th>$\text{rho-4}$</th>
<th>$SD_{\text{rho-4}}$</th>
<th>$\text{rho-5}$</th>
<th>%VE</th>
<th>80% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>28</td>
<td>2830</td>
<td>.43</td>
<td>.15</td>
<td>.63</td>
<td>.12</td>
<td>.83</td>
<td>67%</td>
<td>.47 - .78</td>
</tr>
</tbody>
</table>

The meta-analytical outcomes in Table 7 show a correlation between $d \times \text{aggregated } g$ of .83 and a percentage explained variance of 67.

The value of the correlation calculated with the aggregated $g$ is substantially higher than the value of the correlation calculated with the original $g$ (.83 versus .62,
respectively). We will elaborate on this in the following section.

Table 8

Overview of groups used to compute values of $d$ for the meta-analysis

<table>
<thead>
<tr>
<th>article</th>
<th>test</th>
<th>group</th>
<th>$N$</th>
<th>age</th>
<th>country</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rushton &amp; Skuy (2000)</td>
<td>SPM</td>
<td>Whites</td>
<td>136</td>
<td>17-23</td>
<td>South Africa</td>
</tr>
<tr>
<td>Rushton &amp; Skuy (2000)</td>
<td>SPM</td>
<td>Blacks</td>
<td>173</td>
<td>17-23</td>
<td>South Africa</td>
</tr>
<tr>
<td>Rushton, Skuy, &amp; Fridjhon (2002)</td>
<td>SPM</td>
<td>Whites</td>
<td>86</td>
<td>17-23</td>
<td>South Africa</td>
</tr>
<tr>
<td>Rushton, Skuy, &amp; Fridjhon (2002)</td>
<td>SPM</td>
<td>Indians</td>
<td>58</td>
<td>17-23</td>
<td>South Africa</td>
</tr>
<tr>
<td>Rushton, Skuy, &amp; Fridjhon (2002)</td>
<td>SPM</td>
<td>Blacks</td>
<td>198</td>
<td>17-23</td>
<td>South Africa</td>
</tr>
<tr>
<td>* Rushton (2002)</td>
<td>SPM</td>
<td>Whites</td>
<td>1056</td>
<td>14</td>
<td>South Africa</td>
</tr>
<tr>
<td>* Rushton (2002)</td>
<td>SPM</td>
<td>Indians</td>
<td>1063</td>
<td>14</td>
<td>South Africa</td>
</tr>
<tr>
<td>* Rushton (2002)</td>
<td>SPM</td>
<td>Colored</td>
<td>778</td>
<td>14</td>
<td>South Africa</td>
</tr>
<tr>
<td>* Rushton (2002)</td>
<td>SPM</td>
<td>Blacks</td>
<td>1093</td>
<td>14</td>
<td>South Africa</td>
</tr>
<tr>
<td>Lynn, Allik, &amp; Irving (2004)</td>
<td>SPM</td>
<td>Whites</td>
<td>2735</td>
<td>12-18</td>
<td>Estonia</td>
</tr>
<tr>
<td>Zaaiman, van der Flier, &amp; Thijs (2001)</td>
<td>SPM</td>
<td>Blacks</td>
<td>126</td>
<td>19.5</td>
<td>South Africa</td>
</tr>
<tr>
<td>Rushton, Skuy, &amp; Fridjhon (2003)</td>
<td>APM</td>
<td>Whites</td>
<td>67</td>
<td>17-23</td>
<td>South Africa</td>
</tr>
<tr>
<td>Rushton, Skuy, &amp; Fridjhon (2003)</td>
<td>APM</td>
<td>Indians</td>
<td>40</td>
<td>17-23</td>
<td>South Africa</td>
</tr>
<tr>
<td>Rushton, Skuy, &amp; Bons (2004)</td>
<td>APM</td>
<td>Whites</td>
<td>72</td>
<td>17-23</td>
<td>South Africa</td>
</tr>
<tr>
<td>* Rushton, Skuy, &amp; Bons (2004)</td>
<td>APM</td>
<td>Indians</td>
<td>57</td>
<td>17-23</td>
<td>South Africa</td>
</tr>
<tr>
<td>Rushton, Skuy, &amp; Bons (2004)</td>
<td>APM</td>
<td>Blacks</td>
<td>177</td>
<td>17-23</td>
<td>South Africa</td>
</tr>
<tr>
<td>* Vigneau &amp; Bors (2005)</td>
<td>APM</td>
<td>Whites</td>
<td>506</td>
<td>17-30</td>
<td>Canada</td>
</tr>
<tr>
<td>Zaaiman, van der Flier, &amp; Thijs (2001)</td>
<td>APM</td>
<td>Blacks</td>
<td>126</td>
<td>19.5</td>
<td>South Africa</td>
</tr>
</tbody>
</table>

Note. * Studies marked with an asterisk are excluded as outliers in different stages of the analysis.

Discussion

Spearman’s Hypothesis states that group differences in IQ scores are a function of the cognitive complexity of these IQ scores. In order to answer the question if group differences in intelligence are on the $g$ factor we conducted meta-analyses using data of both Raven’s Standard Progressive Matrices, and Raven’s Advanced Progressive Matrices. Both tests measure the construct of general intelligence quite well and are thus well suited to answer our question. We tested Spearman’s Hypothesis by correlating $g$ loadings of items with the group differences on the items; a high correlation indicates that group differences can be explained by the cognitive
complexity of the items. To test if there are differences we used groups of people from different ethnical backgrounds.

Differences in IQ for individuals within one homogeneous group have been shown to be strongly heritable (Jensen, 1998). This is an important finding, because general intelligence is the key predictor for job- and educational performance. However, there is no consensus to what degree differences between groups are heritable. Te Nijenhuis and Grimen (2007) showed there is a meta-analytical correlation of 1 between $g$ loadings and heritabilities of subtests of an IQ battery. So, a high meta-analytical correlation between $g$ loadings and group differences would strongly suggest there may be a substantial genetic component to group differences.

Based on a large number of studies we found a very strong meta-analytical correlation between $g$ loadings of items and group differences on items. Also, the amount of variance explained in the data points in the meta-analysis is quite high. As a previous meta-analysis has already shown that $g$ loadings and heritabilities are virtually interchangeable, the high correlation between $g$ loadings and group differences in the present meta-analysis is in line with the hypothesis of a substantial genetic component in group differences.

First, we calculated the $g$ loadings based upon individual studies with the $d$ scores. This yielded a correlation, after correcting for statistical artifacts, between $dxg$ of .62. However, when we used an aggregated $g$ based on all relevant data we found a much higher correlation of .83. So, there is a difference in meta-analytical outcomes between using the $g$ loadings from individual studies and the aggregated $g$ loadings. When we take the aggregated $g$ loadings based on a very large total sample as the most optimal estimate, it suggests our estimates based on individual studies are systematically too low.

The correlation between $d$ and our aggregated $g$ of .83 is strong indicating that the difference found in item scores between Blacks, Indians, Roma, and Whites can be strongly explained by the factor of general intelligence. This leaves little room for alternative interpretations, such as cultural bias.

Jensen (1998) has shown that when Black and White children are matched on $g$, Black children outscore White children on subtests of short-term memory and White children outscore Black children on certain spatial subtests. When there are several of these subtests in a test of Spearman’s hypothesis, they can be considered
outliers, and outliers have the well-known effect of a lowering the correlation. Te Nijenhuis and Dragt (2010) reported a meta-analytical correlation of .91, but taking into consideration that many of the datasets in their meta-analysis included subtests measuring short-term memory and specific spatial abilities, the correlation between true \(d\) and true \(g\) without these two outliers is virtually indistinguishable from a correlation of 1.00. So, instead of concluding there is only little room left for cultural explanations of group differences, one could argue there is no room or virtually no room left for cultural explanations.

The same reasoning could be applied to the present psychometric meta-analysis at the item level. Lynn, Allik, and Irwing (2004) convincingly show that the Raven’s, besides \(g\), also measures 1) Gestalt continuation, 2) Verbal-analytical reasoning, and 3) Visio spatial ability. Insofar as there are group differences in the scores on these factors, in tests of Spearman’s hypothesis the relevant items will function as outliers, and will lower the meta-analytical correlation. Follow-up research is necessary to test this hypothesis.

During this project we came across a lot of challenges, mostly of a statistical nature. We consulted renowned researchers in our effort to find the optimal ways to calculate \(g\) loadings. Eventually we came up with a technique, as specified in Method, that we strongly believe is the correct way to calculate these \(g\) loadings. A further statistical issue was how researchers normalized their data. In most of the articles it is simply stated that the data were normalized to standard scores, not specifying how this was done. In our opinion this confusion ends when researchers clearly describe this normalization. The accessibility of this statistically complex research will increase by explaining every single step.

Part of shedding more light on this field of study will have to come from the researchers who have already carried out research on this topic. With websites such as scholar.google.com people have access to a huge database of different studies. For the Raven’s Progressive Matrices alone a great many of articles have been written. However, of course not everyone conducts their analyses in the same way. This has been a huge problem for us when we gathered our data. Since we were dealing with data at the item level we tried to obtain as many original datasets as possible. Luckily some researchers reported the required data at the item level in their article, made them available online, or were kind enough to send them to us. Unfortunately, data from a considerable amount of other studies was not available so studies could not be
Future Research

Further research should include data on Raven’s SPM Plus, which we were unfortunately unable to include in our meta-analysis. As John Raven (personal communication, 2010) pointed out to us, the SPM Plus is superior to the SPM because of its increased difficulty and this ends the ceiling effect present in certain studies using the SPM.

John Raven (2010; see also Wicherts & Johnson, 2009) criticized the approach taken in the present study to test Spearman’s hypothesis at the item level, using analyses based on Classical Test Theory. Raven states: “One simply cannot apply thoughtways derived from Classical Test Theory to measure developed according to the principles of Item Response Theory.” We assume that Raven would argue that the outcomes of a collection of such studies in a psychometric meta-analysis would be uninterpretable. However, the outcomes of a psychometric meta-analysis of the items of the Raven’s are a rho of .83 with 67% of the variance in the data points explained, which is highly similar to the outcomes of a psychometric meta-analysis of IQ batteries: rho is .91 with 98% of the variance in the data points explained. So it would appear that testing Spearman’s hypothesis on a measure developed according to the principles of Item Response Theory and testing Spearman’s Hypothesis on IQ batteries developed according to the principles of Classical Test Theory leads to highly similar conclusions. These findings suggest the analyses in the present study are sound.

Professor Raven (personal communication, 2010) states that he has in his possession many datasets from all over the world. Adding these datasets to the present database would enable carrying out a future, more extensive meta-analysis.

Our fourth specification for inclusion of datasets in the meta-analysis results in the exclusion of the g loadings from a German sample from 1982 (Raven, 1982). Expanding the database might provide more support for the correctness of our choice.

Expanding the database would also make it possible to match groups that are already in our database to comparable new groups. For example, we did not have samples of the same age category to match the twins from Bouchard (Bouchard, Lykken, McGue, Segal, & Tellegen, 1990) and the Mexicans from Lynn, Baehkoff, and Contreras-Niño (2003) with.

A further possibility would be to use the large Estonian dataset from Lynn,
Allik, and Irwing (2004) to create pseudo groups with high IQs and low IQs. These pseudo groups would have the same ethnic background, leading to interesting tests of Spearman’s hypothesis.

**General Conclusion**

Te Nijenhuis and Dragt (2010) found a meta-analytical correlation between $g$ loadings and standardized group differences for test batteries of .91. In the present study, using much less and much smaller studies, and using item scores, we find a correlation of .83. These two values are highly similar and suggest the meta-analytical methodology is basically sound. The outcomes of our meta-analyses strongly support Spearman’s hypothesis. We can conclude that the differences found on test scores between Whites, Indians, Roma, and Blacks can be very strongly attributed to differences in general intelligence.

Based on the results of this paper, the paper by te Nijenhuis and Dragt (2010), and te Nijenhuis and Jongeneel-Grimen (2007), we can conclude that differences between groups on intelligence tests ($d$), general intelligence ($g$), and the heritability coefficient ($h^2$) all correlate highly. This finding is in line with a substantial genetic component in group differences.

**Literature**

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


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, the Netherlands.

hypothesis. Unpublished manuscript, University of Amsterdam, the Netherlands.


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, the Netherlands.

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, the Netherlands.


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


---

1. The paragraphs followed by “(te Nijenhuis & Dragt, 2010)” make verbatim use of te Nijenhuis and Dragt (2010).
2. The first paragraph makes verbatim use of te Nijenhuis and Dragt (2010) with the exception of the numbers that are used.