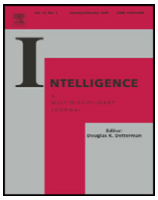
Intelligence 41 (2013) 802–807

Contents lists available at ScienceDirect

Intelligence



Is the Flynn effect on g?: A meta-analysis

Jan te Nijenhuis [a,*], Henk van der Flier [b]

[a] Work and Organizational Psychology, University of Amsterdam, Amsterdam, The Netherlands
[b] Social and Organizational Psychology, Free University, Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Received 30 November 2012
Received in revised form 31 January 2013
Accepted 5 March 2013
Available online 27 March 2013

Keywords:
Flynn effect
Secular score gains
Group differences in intelligence
Method of correlated vectors

ABSTRACT

Black/White differences in mean IQ have been clearly shown to strongly correlate with *g* loadings, so large group differences on subtests of high cognitive complexity and small group differences on subtests of low cognitive complexity. IQ scores have been increasing over the last half century, a phenomenon known as the Flynn effect. Flynn effect gains are predominantly driven by environmental factors. Might these factors also be responsible for group differences in intelligence? The empirical studies on whether the pattern of Flynn effect gains is the same as the pattern of group differences yield conflicting findings. A psychometric meta-analysis on all studies with seven or more subtests reporting correlations between *g* loadings and standardized score gains was carried out, based on 5 papers, yielding 11 data points (total $N = 16{,}663$). It yielded a true correlation of $-.38$, and none of the variance between the studies could be attributed to moderators. It appears that the Flynn effect and group differences have different causes. Suggestions for future research are discussed.

© 2013 Elsevier Inc. All rights reserved.

1. Introduction

A first finding is that IQ scores are the best predictors of a large number of important life outcomes (Jensen, 1998; Schmidt & Hunter, 1998) and that some groups show substantial differences in their mean IQ scores, such as Blacks and Whites in the US, and Europeans and non-Western immigrants in Europe (Hunt, 2011). These group differences have been clearly shown to strongly correlate with *g* loadings, so large group differences on subtests of high cognitive complexity and small group differences on subtests of low cognitive complexity.

A second finding is that IQ scores have been increasing over the last half century, a phenomenon known as the Flynn effect. Flynn effect gains are predominantly driven by environmental factors.

We combine these two findings and ask the question whether the Flynn effect and group differences have the same causes. The empirical studies on whether the pattern of Flynn effect gains is the same as the pattern of group differences yield conflicting findings. We carried out a psychometric meta-analysis on all published studies reporting correlations between *g* loadings and standardized score gains attempting to estimate the true correlation.

1.1. Group differences in IQ

Lynn and Vanhanen (2002, 2012) have shown that there are large group differences in mean IQ scores. In the US, for instance, the Black/White difference is about one S.D. (Jensen, 1998).

The method of correlated vectors is a means of identifying variables that are associated with the *g* 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 those same subtests' relations with the variable in question. When the latter variable is dichotomous the relations are usually calculated in terms of an effect size statistic. When the

[*] Corresponding author at: Gouden Leeuw 746, 1103 KR Amsterdam, the Netherlands.
E-mail address: JanteNijenhuis@planet.nl (J. te Nijenhuis).

0160-2896/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.intell.2013.03.001

latter variable is continuous (or nearly so), the relations are usually calculated in terms of a correlation coefficient (Ashton & Lee, 2005). Jensen (1998, ch.10) quite convincingly argues that it is plausible that the true correlation between *g* loadings and B/W IQ test score differences is close to .90. Virtually all other group differences that have been studied using correlations between *g* loadings and group differences show substantial to high positive correlations (Rushton, Čvorovič, & Bons, 2007; Rushton & Jensen, 2003; te Nijenhuis & van der Flier, 2003). What makes these findings so important is that all studies computing correlations between heritabilities of subtests of an IQ battery and the *g* loadings of these same subtests show substantial to strong values of *r* (Jensen, 1987; Pedersen, Plomin, Nesselroade, & McClearn, 1992; Rijsdijk, Vernon, & Boomsma, 2002; Spitz, 1988).

1.2. The Flynn effect

Many studies have shown that IQ scores have been increasing over the last half century (Flynn, 2012). The average gain on standard broad-spectrum IQ tests between 1930 and 1990 was three IQ points per decade. Recently, however, studies from Scandinavia suggest the secular gains may have stopped in Western, industrialized countries, although the gains are still in progress in Estonia (Must, te Nijenhuis, Must, & van Vianen, 2009). The secular gains are massive and the time period too short for large positive genetic changes in the population, so there is strong consensus that the changes must be largely environmental. There may, however, be a quite modest role for a genetic effect called heterosis, meaning that genetically unrelated parents have children with IQs that are slightly higher than the mean IQ of the general population (see Mingroni, 2007; Woodley, 2011).

Are the strong environmental forces causing the scores over generations to rise the same as the forces causing the group differences? Rushton (1999) showed that secular gains from four Western countries had modest to small negative correlations with *g* loadings. Rushton's (1999) finding has been challenged by Flynn (1999a, 1999b, 2000) and Nisbett (2009) claiming there actually is a substantial positive correlation between secular score gains and *g* loadings. If *g* loadings indeed did correlate highly with both environmental and genetic effects, it would make them useless. Since Rushton's study suggesting secular trends are not related to *g*, various other studies have been carried out (Colom, Juan-Espinosa, & García, 2001; Flynn, 1999a, 1999b, 2000; Must, Must, & Raudik, 2003; te Nijenhuis, in press; te Nijenhuis & van der Flier, 2007; Wicherts et al., 2004) yielding correlations ranging from substantial and negative to large and positive. The present paper aims to reduce the uncertainty regarding the question how strongly the Flynn effect is on the *g* factor by carrying out a psychometric meta-analysis of all published studies on this subject.

2. Method

To test the size of the true correlation between *g* loadings of tests and secular score gains (*d*), we carried out a meta-analysis of all published studies reporting correlations between *g* loading of tests and secular score gains. We identified all studies for the meta-analysis by manual search of Jensen (1998, ch. 10) and the journals Personality and Individual Differences, Intelligence, Psychology in the Schools, Journal of School Psychology, and Journal of Clinical Psychology. Additional search strategies were manual searches at four Dutch Universities, and a computer search of library databases available to us, including ERIC, PsycINFO, MEDLINE, PiCarta, Academic search premier, Web of Science, PubMed, Education-line, SSRN, Cogprints, ROAR and Open DOAR. We used the following keywords: Flynn effect, secular score gains, Jensen effects, and method of correlated vectors. The reference sections of the articles and book chapters obtained were checked, and researchers contributing to this specialist discussion were contacted.

To be included in the present review the following criteria had to be met. First, studies had to report secular gains on well-validated tests. Second, to get reliable correlations between *g* and *d* batteries had to be comprised of at least seven subtests. This choice is based on our experience in a psychometric meta-analysis of the correlation between retest effects and *g* loadings (te Nijenhuis, van Vianen, & van der Flier, 2007), where including datasets with less than seven subtests gave inconsistent results. One could choose to counter with extra strong corrections for unreliability, but we decided to increase reliability by simply dropping the small datasets. There is also a practical consideration that limiting oneself to batteries with, for instance, at least 12 subtests would mean that there are virtually no datasets that could be analyzed with the Method of Correlated Vectors. Application of these inclusion rules yielded five papers with 11 correlations between *g* and *d*.

Psychometric meta-analysis (Hunter & Schmidt, 2004) aims to estimate what the results of studies would have been if all studies had been conducted without methodological limitations or flaws. One of the goals of the present meta-analysis is to have a reliable estimate of the true correlation between *g* loadings and secular score gains (*d*). The collected studies were heterogeneous across various possible moderators.

Standardized secular score gains were computed by dividing the raw score gain by the S.D. of the earlier sample. 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 (Jensen & Weng, 1994). Pearson correlations between the standardized score gains and the *g* loadings were computed.

Psychometric meta-analytical techniques (Hunter & Schmidt, 2004) were applied to the resulting 11 r_{gd}s 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 for. In the present study we corrected for five artifacts that alter the value of outcome measures listed by Hunter and Schmidt (2004): (1) sampling error, (2) reliability of the vector of *g* loadings, (3) reliability of the vector of score gains, (4) restriction of range of *g* loadings, and (5) deviation from perfect construct validity. We present the outcomes step by step.

2.1. 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, so for all comparisons we computed harmonic means for sample size using the formula $4/(1/N_1 + 1/N_2)$.

2.2. Correction for reliability of the vector of g loadings

The value of r_{gd} is 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. The collection of datasets in the present study included no *g* vectors for the same battery from different samples and therefore artifact distributions were based upon other studies reporting *g* vectors for two or more samples, as reported in detail by te Nijenhuis et al. (2007). It appears that *g* vectors are quite reliable, especially when the samples are very large. We summarize the data by reporting the correlations for various ranges of the average sample size of two compared samples. For average sample sizes smaller than 600 te Nijenhuis et al. (2007) report average values of $r = .75$, .78, .86, and .72; for *N* between 601 and 1200 average values of $r = .90$, .93 and .88; for *N* between 1201 and 3000 a value of $r = .92$; and for *N* larger than 3000 an average value of $r = .97$. In the present study, in some cases *g* loadings were taken from test manuals or other high-quality studies – usually with a larger *N* than the study in the meta-analysis – so the *N* of these samples was used to estimate the reliability of the vector of *g* loadings in the study instead of the *N* of the study itself.

2.3. Correction for reliability of the vector of score gains

The value of r_{gd} is attenuated by the reliability of the vector of score gains 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 score gains was estimated using the present datasets and additional datasets, comparing samples that took the same test, and that differed little on background variables.

For large samples – up to a total *N* of 1000 – the reliabilities vary from .63 to .96. Flynn (2000) reports gains using the WISC and the WISC-R for the US and Scotland (total $N = 974$), resulting in $r = .86$. Must et al. (2003) report gains on the Estonian version of the National Intelligence Test for both 12- and 14-year-olds (total $N = 688$), resulting in $r = .92$. Lynn and Hampson (1986) report Japanese data on the Kyoto NX 9-15 Intelligence Test for ten-year-olds from both Kyoto and the surrounding towns (total $N = 737$), resulting in $r = .87$; for eleven-year-olds (total $N = 855$) this results in $r = .82$. They also report data for the seven age groups between 9 and 15 of Japanese children on the Ushijima Intelligence Test (total $N = 2735$). For the adjacent age groups this results in correlations of, $r = .75$, .90, .80, .72, .77, and .81, respectively, with an average total $N = 782$. Colom, Andrés-Pueyo, and Juan-Espinosa (1998) report data for males and females separately on the Spanish Primary Mental Abilities Test (total $N = 882$), resulting in $r = .63$. Jensen (1998, p. 320) reports data on ten- and thirteen-year-olds on the Scottish WISC (total $N = 729$), resulting in $r = .96$. There is a substantial amount of variation in these reliability estimates for large samples. Meta-analysis teaches us that we should expect this variability when combining datasets (Hunter & Schmidt, 2004; Schmidt, 1992) and that finding all studies had highly similar outcomes would be a strong reason for concern.

For even larger samples – a total *N* larger than 1000 – the reliabilities vary from .40 to .91. Flynn (2000) reports secular score gains using the WISC and the WISC-R for West-Germany and Austria (total $N = 4177$), resulting in $r = .76$. Wicherts et al. (2004) report data on the DAT for three different school types, so the adjacent school types can be compared; this results in $r = .70$ (total $N = 2869$) and $r = .40$ (total $N = 2660$), respectively. Colom et al. (2001) report data for males and females separately on the Spanish DAT (total $N = 4177$), resulting in $r = .91$.

Therefore, *d* vectors are quite reliable. This is in line with Rushton (1999) reporting a reliable cluster of score gains for the samples described by Flynn (2000).

2.4. Correction for restriction of range of g loadings

The value of r_{gd} is 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) shows that restriction in *g* loadedness strongly attenuates the correlation between *g* loadings and standardized group differences. Hunter and Schmidt (2004, pp. 37–39) state that the solution to range variation is to define a reference population and express all correlations in terms of that reference population. 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 study population standard deviation by the reference group population standard deviation, that is $u = \text{S.D}_{\text{study}}/\text{S.D}_{\text{ref}}$. As the reference we took the tests that are broadly regarded as exemplary for the measurement of the intelligence domain, namely the various versions of the Wechsler tests for children. te Nijenhuis et al. (2007) report the average standard deviation of *g* loadings of the various Dutch and US versions of the WISC-R and the WISC-III is 0.128. So, the S.D. of *g* loadings of all test batteries was compared to the average S.D. in *g* loadings in the Wechsler tests for children. This resulted in the Dutch GATB having a value of *u* larger than 1.00.

2.5. Correction for deviation from perfect construct validity

The deviation from perfect construct validity in *g* attenuates the value of r_{gd}. 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. So any one limited sample of tests will not yield exactly the same *g* as any other limited 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 value of r_{gd} is attenuated by psychometric sampling error in each of the batteries from which a *g* factor has been extracted.

The more tests and the higher their *g* loadings, the higher the *g* saturation of the composite score. The Wechsler tests have a large number of subtests with quite high *g* loadings

Table 1
Studies of correlations between *g* loadings and gain scores.

Study	Test/tests	*R*	N_{harmonic}
Rushton (1999)			
United States 1	WISC + WISC-R	−.34	245
United States 2	WISC-R + WISC-III	−.40	206
West-Germany	HAWIK + HAWIK-R	−.10	124
Austria	HAWIK	−.14	3636
Must et al. (2003)	NIT	−.29	221
Wicherts et al. (2004)			
	WAIS	.67	288
	DAT	−.15	872
te Nijenhuis (in press)			
US working population	GATB	−.19	1835
Applicant bus drivers 1	GATB	.04	436
Applicant bus drivers 2	GATB	.35	294
te Nijenhuis et al. (2007)	GALO	−.32	3012

Note. Values of statistical measures were computed using information from the best possible sources. When possible, statistical output was taken from the original articles. In some datasets there was no clear Flynn effect – large decreases on a substantial number of subtests –, so they were not included in the meta-analysis.
Data from Rushton (1999). *g* loadings for the WISC were based on the WISC manual (Wechsler, 1949); *g* loadings of the WISC-R were taken from Rushton (1999). *g* loadings from the German WISC were based on data from the manual (Hardesty & Priester, 1956).
Data from Wicherts et al. (2004). WAIS *g* loadings were computed using the reported correlation matrices; DAT (HAVO) *g* loadings taken from te Nijenhuis, Evers, and Mur's (2000) study ($N = 318$) where all nine tests of the DAT battery were employed.

resulting in a highly *g*-saturated composite score. Jensen (1998, pp. 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 a 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 is in the +.80s. When we take this value as an indication of the degree to which an IQ score is a reflection of "true" *g*, we can estimate that a tests' *g* score correlates about .85 with "true" *g*. As *g* loadings are the correlations of tests with the *g* score, it is most likely that most empirical *g* loadings will underestimate "true" *g* loadings; so, empirical *g* loadings correlate about .85 with "true" *g* loadings. As the Schmidt and Le computer program only includes corrections for the first four artifacts the correction for deviation from perfect construct validity was carried out on the value of r_{gd} after correction for the first four artifacts. To limit the risk of overcorrection, we conservatively chose the value of .90 for the correction, yielding a correction factor of 1.11.

3. Results

The results of the studies on the correlation between *g* loadings and secular score gains (*d*) are shown in Table 1. The table gives data derived from eleven studies, with participants numbering a total of 16,663, yielding a harmonic *N* of 11,053. The table gives the reference for the study, the cognitive ability test or tests used, the correlation between *g* loadings and secular score gains, the sample size, and background information on the study. It is clear that virtually all correlations are small to modest.

Table 2 presents the results of the psychometric meta-analysis of the eleven data points. It shows (from left to right): the number of correlation coefficients (*K*), total sample size based on harmonic means (N_h), the mean observed correlations (*r*) and their standard deviation (S.D.r), the true correlations one can expect once artifactual error from unreliability in the *g* vector and the *d* vector and range restriction in the *g* vector has been removed (rho), and their standard deviation (S.D.rho). The next two columns present the percentage of variance explained by artifactual errors (%VE) and the 80% credibility interval (80% CI). This interval denotes the values one can expect for rho in sixteen out of twenty cases.

The estimated true correlation has a value of −.26, but only 7% of the variance in the observed correlations is explained by artifactual errors. However, Hunter and Schmidt (2004) state that extreme outliers should be left out of the analyses, because they are most likely the result of errors in the data. They also argue that strong outliers artificially inflate the S.D. of effect sizes and thereby reduce the amount of variance that artifacts can explain. We chose to leave out two outliers comprising 4% of the research participants. This resulted in a small change in the value of the true correlation, a large decrease in the S.D. of the observed *r* with 49% and a very large decrease in the S.D. of rho with 100%, and a very large increase in the amount of variance explained in the observed correlations by statistical artifacts. So, when the two outliers are excluded artifacts explain all of the variance in the observed correlations. Finally, a correction for deviation from perfect construct validity in *g* took place, using a conservative value of .90, yielding a correction factor of 1.11. This resulted in a value of −.38 for the final estimated true correlation between *g* loadings and secular score gains.

Table 2
Meta-analysis results for correlations between *g* loadings and secular gain scores after corrections for 1) sampling error, 2) reliability of the vector of *g* loadings, 3) reliability of the vector of score gains, 4) restriction of range of *g* loadings, and 5) deviation from perfect construct validity.

Studies included and corrections	*K*	N_h	*r*	S.D.r	rho	S.D.rho	%VE	80% CI
All	11	11,053	−.17	.191	−.26	.276	7	−.62 to .09
All minus 2 outliers								
1	9	10,587	−.21	.097	−.21	.093	8	−.32 to −.09
1 + 2	9	10,587	−.21	.097	−.21	.095	9	−.33 to −.09
1 + 2 + 3	9	10,587	−.21	.097	−.25	.108	17	−.39 to −.12
1 + 2 + 3 + 4	9	10,587	−.21	.097	−.34	.000	111	−.34 to −.34
1 + 2 + 3 + 4 + 5	9	10,587	−.21	.097	−.38	.000	111	−.34 to −.34

Note. K = number of correlations; N_h = total sample size based on harmonic means (with a total $N = 16{,}663$); r = mean observed correlation (sample-size weighted); S.D.r = standard deviation of observed correlation; rho = true correlation (observed correlation corrected for unreliability and range restriction); S.D.rho = standard deviation of true correlation; %VE = percentage of variance accounted for by artifactual errors; 80% CI = 80% credibility interval.

The outcome of any meta-analysis based on a limited number of studies depends to some extent on study properties that vary randomly across studies. This phenomenon is called "second-order sampling error"; the error stems from the sampling of studies in a meta-analysis. Percentages of variance explained larger than 100% are not uncommon with a limited number of studies. The correct conclusion is that all the variance is explained by statistical artifacts (see Hunter & Schmidt, 2004, pp. 399-401 for an extensive discussion).

4. Discussion

Black/White differences are strongly linked to *g* loadings. Average scores on intelligence tests have been rising substantially and consistently, all over the world, and are predominantly driven by environmental factors. Are these factors also responsible for group differences in intelligence? Is the pattern of secular score gains the same as the pattern of group differences?

We reduced the uncertainty regarding the question how strongly the Flynn effect is on the *g* factor by carrying out a psychometric meta-analysis on all studies reporting correlations between *g* loadings and score gains. A psychometric meta-analysis based on a large total $N = 16{,}663$ shows that after corrections for several statistical artifacts there is an estimated true correlation of $-.38$ between *g* loadings of tests and secular score gains. Jensen estimated that the true correlation between *g* loadings and Black/White IQ test score differences is close to .90. So, secular score gains and group differences have highly different correlations with *g* loadings, which suggests they have different causes. The Flynn effect is predominantly caused by environmental factors, and it is less plausible that these same environmental factors play an important role in explaining group differences in IQ scores.

There are strong differences of opinion on the causes of group differences in IQ: Rushton and Jensen (2005, 2010) argue that there is a strong genetic component to group differences, whereas Nisbett (2009) argues that group differences are wholly due to Blacks and non-Western immigrants growing up in a lower-quality environment (see also Dawkins, 1982; Meisenberg, 2010; Nyborg, 2012, Wicherts, Dolan, Carlson, & Maas, 2010). Rindermann, Woodley, and Stratford (2012) show, based on evolutionary theory, that there are substantial correlations between mean IQ scores of countries and genetic markers.

The position could be taken that one should not expect a single true correlation from Flynn effect studies that involve different age groups, different populations, and different time periods. For example, suppose nutrition played a more important role in the secular gains observed between the Great Depression and the post-WWII era, but contributed relatively little to gains observed between 1960 and 1980. Instead, suppose most of the gains between the 1960s and 1980s were due to practice effects. In this scenario the correlation between *g* loadings and Flynn effects would vary among studies. However, there is simply no support at all for these moderator variables: all the variance between the studies is explained by four statistical artifacts, namely sampling error, reliability of the *g* vector, reliability of the *d* vector, and restriction of range. As several artifacts explain all the variance in the effect sizes, other dimensions on which the studies differ play no role of significance.

The two outliers excluded are statistical outliers: we have no good explanation why their values of r_{dg} are so different from the values from the other nine data points, which are perfectly homogeneous. The Flynn effects differ dramatically by narrow ability: Flynn (2007, p. 8) shows huge gains on the Wechsler subtest Similarities and small gains on the subtests Information, Arithmetic, and Vocabulary. One could argue that it is the collection of narrow abilities in a battery that is the key, because r_{dg} is computed on a small number of observations, and is therefore highly sensitive to a small number of weak or strong effects in the data. So, having an unusual amount of subtests like Similarities, or having an unusual amount of subtests like Information, Arithmetic, and Vocabulary may, for instance, yield a strong positive correlation instead of the common modest negative correlation. However, one of the outliers is found for the GATB, whereas three other data points based on the GATB fit perfectly into the normal pattern. Moreover, the other outlier is found for the Dutch WAIS, whereas four other data points based on various other versions of the Wechsler also fit perfectly into the normal pattern. This puzzle remains to be solved.

The present meta-analysis is based on eleven data points from studies that explicitly test the correlation between gains and *g* loadings, but there are studies in the Flynn effect literature that do not report $r(g \times d)$. It would be a good idea to carry out half a dozen or a dozen new studies into the correlation between *g* and secular gains, based on at least seven subtests, and add these data points to a new more extensive meta-analysis. Although we found no support at all for moderators in the present meta-analysis, it is possible that a larger database may yield some support for region of the world or period as moderators. Moreover, the fact that there was a modest amount of second-order sampling error also suggests the need for some additional studies for a future, extended meta-analysis.

[*]References

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

Colom, R., Andrés-Pueyo, A., & Juan-Espinosa, M. (1998). Generational IQ gains: Spanish data. *Personality and Individual Differences*, *25*, 927–935.

Colom, R., Juan-Espinosa, M., & García, L. F. (2001). The secular increase in test scores is a "Jensen effect". *Personality and Individual Differences*, *30*, 553–559.

Dawkins, R. (1982). *The extended phenotype. The long reach of the gene.* Oxford: Oxford University Press.

Flynn, J. R. (1999a). Evidence against Rushton: The genetic loading of WISC-R subtests and the causes of between-group IQ differences. *Personality and Individual Differences*, *26*, 373–379.

Flynn, J. R. (1999b). Reply to Rushton. A gang of gs overpowers factor analysis. *Personality and Individual Differences*, *26*, 391–393.

Flynn, J. R. (2000). IQ gains, WISC subtests and fluid g: g theory and the relevance of Spearman's hypothesis to race. In G. R. B. J. Goode (Ed.), *The nature of intelligence* (pp. 202–227). New York: Wiley.

Flynn, J. R. (2007). *What is intelligence?* : Cambridge University Press.

Flynn, J. R. (2012). *Are we getting smarter? Rising IQ in the twenty-first century.* Cambridge University Press.

Hardesty, F. P., & Priester, H. J. (1956). *Handbuch für den Hamburg–Wechsler Intelligenztest für Kinder [Manual of the Hamburg Wechsler Intelligence Test for Children].* Germany, Stuttgart: Huber.

Hunt, E. (2011). *Human intelligence.* Cambridge University Press.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). London: Sage.

Jensen, A. R. (1987). Individual differences in mental ability. In J. A. Glover, & R. R. Ronning (Eds.), *Historical foundations of educational psychology*. New York: Plenum.

[*] References marked with an asterisk indicate studies included in the meta-analysis.

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

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

Lynn, R., & Hampson, S. (1986). The rise of national intelligence: Evidence from Britain, Japan and the U.S.A. *Personality and Individual Differences*, *7*, 23–32.

Lynn, R., & Vanhanen, T. (2002). *IQ and the wealth of nations*. London: Praeger.

Lynn, R., & Vanhanen, T. (2012). *Intelligence: A unifying construct for the social sciences*. London: Ulster Institute for Social Research.

Meisenberg, G. (2010). The reproduction of intelligence. *Intelligence*, *38*, 220–230.

Mingroni, M. A. (2007). Resolving the IQ paradox: Heterosis as a cause of the Flynn effect and other trends. *Psychological Review*, *114*, 806–829.

*Must, O., Must, A., & Raudik, V. (2003). The secular rise in IQs: In Estonia the Flynn effect is not a Jensen effect. *Intelligence*, *167*, 1–11.

Must, O., te Nijenhuis, J., Must, A., & van Vianen, A. E. M. (2009). Comparability of IQ scores over time. *Intelligence*, *37*, 25–33.

Nisbett, R. E. (2009). *Intelligence and how to get it*. New York: Norton.

Nyborg, H. (2012). The decay of Western civilization: Double relaxed Darwinian Selection. *Personality and Individual Differences*, *53*, 118–125.

Pedersen, N. L., Plomin, R., Nesselroade, J. R., & McClearn, G. E. (1992). A quantitative genetic analysis of cognitive abilities during the second half of the life span. *Psychological Science*, *3*, 346–353.

Rijsdijk, F. V., Vernon, P. A., & Boomsma, D. I. (2002). Application of hierarchical genetic models to Raven and WAIS subtests: A Dutch twin study. *Behavior Genetics*, *32*, 199–210.

Rindermann, H., Woodley, M. A., & Stratford, J. (2012). Haplogroups as evolutionary markers of cognitive ability. *Intelligence*, *40*, 362–375.

*Rushton, J. P. (1999). Secular gains in IQ not related to the g factor and inbreeding depression — unlike Black–White differences: A reply to Flynn. *Personality and Individual Differences*, *26*, 381–389.

Rushton, J. P., Čvorovič, 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., & Jensen, 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*, *2003*, 177–183.

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

Rushton, J. P., & Jensen, A. R. (2010). Editorial. The rise and fall of the Flynn effect as a reason to expect a narrowing of the Black–White IQ gap. *Intelligence*, *38*, 213–219.

Schmidt, F. L. (1992). What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. *American Psychologist*, *47*, 1173–1181.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, *124*, 262–274.

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

Spitz, H. H. (1988). Wechsler subtest patterns of mentally retarded groups: Relationship to *g* and to estimates of heritability. *Intelligence*, *12*, 279–297.

*te Nijenhuis, J. (in press). The Flynn effect, group differences, and g loadings. Personality and Individual Differences. http://dx.doi.org/10.1016/j.paid.2011.12.023.

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. (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. (2007). The secular rise in IQs in the Netherlands: Is the Flynn effect on *g*? *Personality and Individual Differences*, *43*, 1259–1265.

te Nijenhuis, J., van Vianen, A. E. M., & van der Flier, H. (2007). Score gains on *g*-loaded tests: No *g*. *Intelligence*, *35*, 283–300.

Wechsler, D. (1949). *Manual for the Wechsler Intelligence Scale for Children (WISC)*. New York: The Psychological Corporation.

Wicherts, J. M., Dolan, C. V., Carlson, J. S., & Maas, H. L. J. v. d. (2010). Raven's test performance of sub-Saharan Africans: Average performance, psychometric properties, and the Flynn Effect. *Learning and Individual Differences*, *20*, 135–151.

*Wicherts, J. M., Dolan, C. V., Oosterveld, P., van Baal, G. C. V., Boomsma, D. I., & Span, M. M. (2004). Are intelligence tests measurement invariant over time? Investigating the nature of the Flynn effect. *Intelligence*, *32*, 509–537.

Woodley, M. A. (2011). Heterosis doesn't cause the Flynn effect: A critical examination of Mingroni (2007). *Psychological Review*, *118*, 689–693.