Intelligence 43 (2014) 27–34

Contents lists available at ScienceDirect

Intelligence

Controlling for increased guessing enhances the independence of the Flynn effect from *g*: The return of the Brand effect

Michael Anthony Woodley [a,b,*], Jan te Nijenhuis [c], Olev Must [d], Aasa Must [e]

[a] Umeå University, Department of Psychology, Umeå, Sweden
[b] Center Leo Apostel for Interdisciplinary Research, Vrije Universiteit Brussel, Belgium
[c] University of Amsterdam, Work and Organizational Psychology, The Netherlands
[d] University of Tartu, Department of Psychology, Estonia
[e] Estonian National Defence College, Estonia

ARTICLE INFO

Article history:
Received 26 June 2013
Received in revised form 4 November 2013
Accepted 16 December 2013
Available online xxxx

Keywords:
Brand effect
Flynn effect
g loadings
Jensen effect

ABSTRACT

The cause of the Flynn effect is one of the biggest puzzles in intelligence research. In this study we test the hypothesis that the effect may be even more independent from *g* than previously thought. This is due to the fact that secular gains in IQ result from at least two sources. First, an authentic Flynn effect that results from environmental improvements and should therefore be strongly negatively related to the *g* loading (and therefore the heritability) of IQ subtests. Second, a "Brand effect", which results from an increase in the number of correct answers simply via enhanced guessing. As harder items should encourage more guessing, secular gains in IQ stemming from this Brand effect should be positively associated with subtest *g* loadings. Analysis of Estonian National Intelligence Test data collected between 1933 and 2006, which includes data on guessing, *g* loadings and secular IQ gains, corroborates this hypothesis. The correlation between gains via the Brand effect and *g* loadings is .95, as predicted. There is a modest negative association between raw secular gain magnitude and subtest *g* loadings (−.18) that increases to −.47 when these are controlled for the Brand effect. Applying five psychometric meta-analytic corrections to this estimate raises it to −.82 indicating that the authentic Flynn effect is substantially more independent from *g* than previously thought.

1. Introduction

Much controversy surrounds the finding of a three point per decade secular gain in measured IQ (Flynn, 1984, 1987, 2009, 2012), usually referred to as the Flynn effect (Herrnstein & Murray, 1994). A variety of factors have been proposed as causative of this effect. These include nutrition (Flynn, 1987), education (Husén & Tuijnman, 1991), improvements in hygiene (Eppig, Fincher, & Thornhill, 2010), decreases in environmental neurotoxin levels (Nevin, 2000), increased familiarity with or sensitivity to the solution rules of tests (Armstrong & Woodley, 2014) and the presence of cultural amplifiers, which via positive feedback lead to large gains in IQ on the basis that smarter populations also demand greater cognitive stimulation (Dickens & Flynn, 2001). Heterosis or hybrid vigor has also been proposed as both a minor and a major causative factor in secular gains (Jensen, 1998a; Mingroni, 2004, 2007). Another theory is that the gains result from changing test-taking habits — specifically the tendency towards the use of rapid guessing on timed multiple-choice-type answer formats under circumstances where easily learned strategies can be used to reduce the numbers of wrong answers, thus increasing the odds of selecting correct answers by chance alone (Brand, 1987a,b, 1990, 1996; Brand, Freshwater, & Dockrell, 1987). Jensen (1998a) argued that the rapidity of the gains rules out a non-environmental origin except possibly in the US during the opening decades of the 20th century, where increased admixture with European immigrants might have resulted in small gains due to heterosis.

* Corresponding author at: Umeå University, Department of Psychology, Umeå, Sweden.
E-mail address: Woodley@psy.umu.se (M.A. Woodley).

0160-2896/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.intell.2013.12.004

1.1. Gains in g?

One of the biggest controversies surrounding the potential causes of the gains concerns the idea that they may be in some way associated with a change in the level of *g* or the general factor of intelligence within populations. Some have argued that this must be the case as the effect seems to be associated with real-world improvements such as increasing precociousness in games like chess, bridge and go (Howard, 2001) and increasing brain size (Lynn, 1989). Much progress has been made towards better understanding the causes of the gains via the use of the method of correlated vectors, or the correlation between the *g* loading of a subtest and the size of the gains associated with that subtest. A strong positive vector correlation between secular gains and subtest *g* loading would indicate that the gains are a Jensen effect i.e. that their relation with intelligence is positively mediated by *g*. An anti-Jensen effect means that there is a strong negative correlation between the *g* vector and the *d* vector and strongly suggests that the effect is independent of *g*, and that the effect instead occurs on subtest-specific sources of variance. Different studies have produced different results from applying this method to the pattern of secular gains. These results range from the finding that with respect to literacy in Estonia the gains are a perfect anti-Jensen effect (−1; Must, Must, & Raudik, 2003a), to the finding that they are a strong Jensen effect on Fluid intelligence measures in Spain (.78; Colom, Juan-Espinosa, & García, 2001). The preponderance of studies do however indicate that the pattern of gains are either negatively or non-correlated with subtest *g* loadings. This was demonstrated recently in a psychometric meta-analysis of over 17,000 individuals and 12 studies, which revealed that the 'true' vector correlation between the pattern of secular gains and subtest *g* loadings is −.38 (te Nijenhuis & van der Flier, 2013). This indicates that the gains are definitely not a Jensen effect and are substantially but perhaps not completely independent of *g*. The meta-analysis also shows that the differences in outcomes between various studies can be perfectly explained by just five statistical artifacts, such as sampling error and reliability. When the theoretically expected effect is +1 or −1 the method of correlated vectors can most likely be applied to test batteries with as few as four subtests. However, when the theoretically expected effect is in the vicinity of 0 the method appears to become very sensitive to the effect of outliers and most likely requires at least seven subtests for a reliable outcome. This could explain the extreme outlier that is the study by Colom et al. (2001), as it was based on just five subtests.

The issue of whether or not the pattern of secular gains is a Jensen effect is important in terms of inferring causation. Wholly genetically-influenced variables, such as subtest heritabilities (Rushton & Jensen, 2010), inbreeding depression (Jensen, 1998a; Rushton, 1999) and hybrid vigor or heterosis (Nagoshi & Johnson, 1986) are associated with strong Jensen effects in all cases, whereas purely environmental effects, such as IQ gains via retesting (te Nijenhuis, van Vianen, & van der Flier, 2007) and gains due to adoption (Jensen, 1998b) are strong anti-Jensen effects. So, there is a cluster of genetic effects yielding a correlation of +1 with *g* loadings and a cluster of cultural–environmental effects yielding a correlation of −1 with *g* loadings. On this basis, certain causal theories of the Flynn effect can be ruled out, such as the idea that it results primarily from the effects of heterosis (Mingroni, 2004, 2007).

As was mentioned previously, on the basis of the results of meta-analysis, the pattern of secular gains is clearly not a Jensen effect. However it is not enough of an anti-Jensen effect to completely rule out potential genetic causes also. Rushton (1999) for example found a clear negative vector correlation between five secular gains and subtest *g* loadings. He also found that four out of the five sets of secular gains included in his analysis exhibited both strong loadings on an environmental factor in a factor analysis and small positive loadings on a genetic factor. te Nijenhuis and van der Flier (2013) argue that there may be a quite modest role for heterosis in secular gains. A quite modest role could be interpreted as 5 to 10% of the overall gains. There are several problems with this hypothesis however, chief amongst which is the fact that inbreeding was never that prevalent in the West historically (Flynn, 2009). Furthermore recent research reveals that levels of *g* have been declining in the West, as indicated by a psychometric meta-analysis of the secular slowing of simple reaction time means between the 19th and 21st centuries (Silverman, 2010; Woodley, te Nijenhuis, & Murphy, 2013). The average decline in *g* across cohorts may be equivalent to around −1.16 points per decade, or −13.35 points between 1889 and 2004. The most likely cause of this is the presence of dysgenic fertility in many Western cohorts between the end of the 19th century and the present day (Woodley et al., 2013). It must be noted that the finding has not been received uncritically (Dodonova & Dodonova, 2013; Flynn, 2013; Nettelbeck, 2014; Silverman, 2013). Despite this, granting our premise, the magnitude of dysgenic fertility is strongly positively mediated by the *g* saturation of subtests, hence is an undisputed Jensen effect (Reeve, Lyerly, & Peach, 2013; Woodley & Meisenberg, 2013). Therefore as *g* cannot be simultaneously rising and falling (Woodley, 2011) an alternative explanation must be sought for both the lack of a perfect anti-Jensen effect on the pattern of secular gains, and the presence of cross-loadings in Rushton (1999).

Finally, a position maintained by some researchers (i.e. Jensen, 1998a) is that for secular gains to be meaningful they must involve gains in *g*, as it is asserted that this is the sole source of criterion validity in IQ tests. The idea therefore is that a 'hollow' secular gain is a meaningless one. This model is increasingly at odds with the data indicating that completely 'hollow' test-score variance is able to predict real-world performance, albeit within narrow parameters (Coyle & Pillow, 2008), and also that the presence of large numbers of individuals capable of cognitively specializing can lead to group-level increases in aggregate efficiency or a sort that might have driven the massive growth in wealth throughout the 20th century (the historical trend in wealth growth strongly parallels the growth in secular gains; Woodley, 2012). A wholly 'hollow' gain in ability strengthens theoretical models requiring that for massive secular gains to have had an impact on the real world, they need to be completely independent of *g* (Flynn, 2009; Woodley, 2012).

1.2. Higher scores due to guessing: the Brand effect

One possibility concerns an older causal theory of secular IQ gains which was proposed by Brand (1996), and is based

on the idea that guessing the answers to questions exhibiting certain timed, multiple-choice answer formats will boost IQ scores simply because easily learned strategies can be used in eliminating obviously wrong answers, hence guessing the answers to questions increases purely by chance the tendency to generate correct answers as more items are attempted, but requires that accuracy be sacrificed. That this mechanism may in fact be important to understanding the pattern of secular gains was demonstrated recently by Must and Must (2013), who estimated that on the subtests of the Estonian National Intelligence Test around a third of the secular gains experienced by the 2006 relative to the 1933/36 Estonian student cohort could be accounted for by the fact that the rate of guessing had increased in the latter cohort. Another recent study by Pietschnig, Tran, and Voracek (2013) similarly found tentative evidence for a contribution to secular gains in IQ stemming from increased guessing behavior amongst an Austrian cohort, using item response theory.

Common sense suggests that easy and less *g*-loaded subtests will lead to little guessing and that difficult, and therefore more *g* loaded subtests will lead to more guessing. Jensen (1998a) argued that *g* loadings (as an index of the complexity of tests) and the difficulty of tests are not strictly equivalent. Arend et al. (2003) found that difficulty (measured by the rate at which participants failed a given item) and complexity (as indicated via item-level *g*-loadings) were correlated at .52, suggesting some construct independence. However, Arend et al. (2003) could be taken to indicate that the two ratings are in fact actually highly correlated. This is because it is generally known that item scores are notoriously unreliable (e.g. Rushton & Skuy, 2000). In Arend et al. (2003) both difficulty and cognitive complexity are measured at the item level, thus doubling the amount of unreliability. This can be illustrated as follows: let's suppose that the reliability of these item measures is .60, which may even be something of an overestimate. Correcting the observed correlation for the unreliability present in both measures yields a corrected correlation of $.52 / (.75 \times .75) = .93$. Thus indicating that in Arend et al. (2003) the measures become almost interchangeable once unreliability is properly controlled. Other research confirms the strong association between complexity and difficulty, for example Helms, van de Vijver, and Poortinga (2003), who tested Spearman's hypothesis on Dutch and immigrant children with both *g* loadings and high-quality measures of cognitive complexity, found that both loaded very strongly on the same factor in a factor analysis. We therefore believe that it is reasonable to treat the two as essentially interchangeable — although in a subsequent analysis we will demonstrate that this is the case in our samples also.

If the degree to which guessing is employed is indeed a function of the difficulty of a subtest (i.e. as reflected in its *g* loading) IQ gains via this Brand effect will pose as a Jensen effect, but *actually* give a 100% *g* independent boost to IQ scores. This is in contrast to purely environmentally driven gains in IQ, which may reflect a genuine increase in performance on narrow sources of ability variance in response to the presence of various intelligence-enhancing social and environmental multipliers (Dickens & Flynn, 2001). In this scenario correcting the observed secular gains for the Brand effect should increase the independence of the gains with respect to *g*, thus unmasking the true magnitude of the anti-Jensen effect on the almost perfectly environmentally-driven *authentic* Flynn effect. Furthermore, the contribution of Brand effects to the pattern of secular gains would also account for Rushton's (1999) aforementioned finding of small positive cross-loadings of secular gains on a genetic factor in factor analysis. What Rushton discovered might therefore have been a spurious association with *g* stemming from the Brand effect in his sample.

Here we test the Estonian data collected by the Musts and others for evidence of the hypothesis that the independence of the Flynn effect from *g* is even stronger than the te Nijenhuis and van der Flier meta-analysis shows because it is masked by the presence of the Brand effect.

2. Methods

We conducted an initial analysis into the relationship between subtest difficulty (as measured using the proportion of wrong answers per subtest) and complexity (as measured using the *g* loadings of the items). We also examine the relationship between these variables and the false positive rate, which is calculated as the percentage difference between the proportion of correct answers and the proportion of correct answers adjusted for guessing (i.e. the number of right answers minus the number of wrong answers based on the formula given in Must & Must (2013)). This analysis is conducted for both the 1933/36 and 2006 cohorts. Separate data on the *g* loadings of the NIT subtests for each cohort were obtained from Must, te Nijenhuis, Must, and van Vianen (2009) who employed confirmatory factor analysis in estimating the values. Whilst we feel that the cognitive complexity/difficulty association is quite robust (as was discussed previously), based on reviewer comments, a demonstration that this was also the case in our samples was deemed necessary.

We then separated the Brand effect from the Flynn effect in the comparison of the differences between the Estonian 1933/36 and 2006 cohorts. Must and Must report the secular gain in the number of right answers and the secular gain in the number of right answers adjusted for guessing per subtest in their Fig. 1. The values for the Brand effect can then simply be computed by taking the secular gain in the number of right answers and subtracting the secular gain in the number of right answers adjusted for guessing. The value for the Brand effect then simply reflects the effect of false positive answers i.e. the secular gain in IQ due to the secular increase in guessing. Data on these variables are available on nine of the 10 subtests in the Estonian National Intelligence Test (NIT; Arithmetic reasoning, Sentence completion, Same–different, Symbol–digit, Computation, Information, Vocabulary, Analogies and Comparisons), with a combined *N* of 1732 individuals (890 in the 1933 cohort and 913 in the 2006 cohort) and a mean age of 13.5 in both cohorts. Data on the Flynn effect were taken from Must and Must (2013) and the 1933/36 *g* loadings for the nine NIT subtests were obtained from Must et al. (2003b), who used principal component analysis (PCA). The PCA derived estimates are used in preference to the CFA derived ones as PCA is an exploratory rather than a model-based procedure, hence there are fewer assumptions that go into the PCA derived estimates.

2.1. Criticism and defense of the method of correlated vectors: the psychometric meta-analytic-MCV hybrid model

We choose to analyze all of our data with the method of correlated vectors (MCV). The MCV has been criticized by various researchers (e.g. Ashton & Lee, 2005; Dolan, 2000; Hunt, 2011, p. 363–365), and, notwithstanding its frequent use, is considered controversial by some. We therefore take this opportunity to defend the method.

Most of the criticism of the MCV rests on two problematic premises. First, Jensen (1998a, pp. 372–374) clearly states that fairly representative samples should be used, a large enough number of tests should be used, and these tests must also be diverse in terms of content. Ashton and Lee show that analyses involving unbalanced collections of tests yield outcomes that make little sense, but Jensen explicitly warned researchers about the use of unbalanced samples. Second, Jensen (1998a, pp. 380–383) shows that there are four statistical artifacts that strongly attenuate the outcomes of the MCV, such as unreliability and restriction of range. Hence, Jensen was well aware that there were weaknesses in his method and he showed that controlling for them strongly increased the value of the resulting vector correlations. Dolan (2000) shows that small samples can yield unreliable outcomes, and this is simply consistent with Jensen's previous statements. Therefore, there is little in these criticisms that Jensen did not anticipate years before.

Indeed, the MCV is not a strong statistic when used in isolation, but in combination with the highly powerful methodology of psychometric meta-analysis (Hunter & Schmidt, 2004) it has the potential to lead to a robust statistic, yielding strong, highly stable meta-analytical outcomes (see: te Nijenhuis & van der Flier, 2013; te Nijenhuis et al., 2007). The advantages of combining the MCV with psychometric meta-analysis are fourfold. First, 100% of the published datasets can be used, including quite small ones. Second, small studies sometimes report *g* loadings based on a small *N*, so the *g* vectors are quite unreliable, but they can simply be substituted by *g* loadings from high-quality samples, such as the large nationally representative samples reported in most test manuals, thereby strongly reducing the unreliability. Third, there is information on the variance between studies, which is generally large. It is important to see how the strong impact of various statistical artifacts leads to highly different outcomes when using otherwise quite comparable studies. Therefore, one application of the MCV yielding an unexpectedly low correlation does not necessarily mean there is a strong lack of support for the theory, but that the data point might instead simply be at the outer edge of the reliability distribution, but not yet an outlier. Critics may therefore erroneously conclude that the MCV is a flawed method, based on the use of such a non-representative dataset. In the meta-analysis on test–retest IQ gains (te Nijenhuis et al., 2007) 99% of the variance between studies was explained, and in the meta-analysis of Flynn effect gains (te Nijenhuis, & van der Flier, 2013) 100% of the variance was explained. Fourth, corrections for several important statistical artifacts can be carried out, leading to a less restricted view of relations between constructs.

Hunt (2011, p. 365) and Dolan (2000) conclude that the MVC is flawed and advise the use of multigroup confirmatory factor analysis (MGCFA) instead. However, when using MGCFA instead of the combination of MCV and psychometric meta-analysis all the aforementioned advantages disappear. First, MGCFA can only be carried out on quite large samples, so the many small datasets simply cannot be analyzed, hence the information contained in them is lost for the purposes of accumulation. In many fields only small-scale experiments can be carried out, based on *N*s that are simply too small for the use of MGCFA. Also, at least the correlation matrices have to be available, therefore many datasets can simply not be analyzed. Of the studies used in the two meta-analyses by te Nijenhuis and his co-authors the large majority of studies could simply not be analyzed with MGCFA, leading to a potentially enormous waste of scientific data. Second, *g* loadings from huge samples are better than the *g* loadings from small or medium-sized samples, but MGCFA as described by Dolan does not import better *g* loadings from other samples. Third, because the focus is on individual datasets there is no information on the variance between studies, and the meta-analyses of te Nijenhuis and co-authors show that there is a massive amount of variance between studies, and that it is essential to understand it and to properly account for it. Fourth, in published studies Dolan does not use the corrections for statistical artifacts common in PMA. So, in all likelihood the outcomes from single studies using MGCFA will in many cases differ dramatically from the outcomes based on the combination of MCV and PMA. MGCFA may well have fundamental flaws and we should begin to discuss whether it ought to be used in the study of group differences or maybe should be drastically revised. It is already a fact that more studies using real-world data based on MCV have been published than studies based on MGCFA; also, quite a few studies on MGCFA are forced to rely on computer-generated data instead of real data.

In an influential article Frank Schmidt (1992) states that meta-analysis changes the way we do research in many ways. Meta-analysis has made clear that the amount of information contained in one study is only modest, and therefore an individual study must be considered only a single data point to be incorporated into a future meta-analysis. Only conclusions based on the huge amount of information in meta-analyses have enough empirical strength to be taken seriously. So, the present study fits into Schmidt's strategy: we carry out the first study of this kind and hope that it will stimulate others to carry out comparable studies, which then can later be meta-analyzed.

The situation with the MCV looks very much like the situation in personnel selection predicting job performance with IQ tests before the advent of meta-analysis. Predictive validities for the same job from different studies were yielding highly variable outcomes and it was widely believed that every new situation required a new validation study. Schmidt and Hunter (1977) however showed that because most of the samples were quite small, there was a massive amount of sampling error. Correcting for this statistical artifact and a small number of others led to an almost complete disappearance of the large variance between the studies in many meta-analyses. The outcomes based on a large number of studies all of a sudden became crystal clear and started making theoretical sense (Gottfredson, 1997). This was a true paradigm shift in selection psychology. Analyzing many studies with MCV and meta-analyzing these studies has already led to clear outcomes and has the potential to lead to improvements in theory within the field of intelligence

research. In an editorial published in *Intelligence*, Schmidt and Hunter (1999) have argued the need for more psychometric meta-analyses within the field.

2.2. The use of the psychometric meta-analytic-MCV hybrid model in the present study

The method of correlated vectors was employed in order to determine the relationships between *g* loadings and i) the secular gains, ii) the secular gains corrected for the secular gain in guessing and iii) the Brand effect, or the effect of false positive answers. The sample sizes for the two groups are highly similar, so we used the mean $N = 902$ for the combination of the two samples, which is used in establishing the significance of the vector correlations, that are in turn computed as Pearson's product moment correlations.

Te Nijenhuis and van der Flier (2013) carried out a psychometric meta-analysis of the correlations between *g* loadings and secular score gains. As was mentioned in the previous section, 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. Following te Nijenhuis and van der Flier, in the present study we corrected the correlation between *g* loadings and gains corrected for guessing for four artifacts that alter the value of outcomes: (1) reliability of the vector of *g* loadings, (2) reliability of the vector of score gains, (3) restriction of range of *g* loadings, and (4) deviation from perfect construct validity. Our description of the methods and the estimates used derive from the te Nijenhuis and van der Flier study.

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. te Nijenhuis et al. (2007) report *g* vectors for two or more samples. It appears that *g* vectors are quite reliable, especially when the samples are very large. te Nijenhuis and van der Flier report for samples with an *N* that is similar to the mean $N = 902$ for the two Estonian samples, and reliabilities of .90, .93, and .88. We use the mean value of .90, which yields a correction factor of 1.05.

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. te Nijenhuis and van der Flier estimated the reliability of the vector of score gains using the present datasets and additional datasets, comparing samples that took the same test, and that differed little on background variables. te Nijenhuis and van der Flier report for samples with an *N* that is similar to the mean $N = 902$ for the two Estonian samples, 12 reliabilities. We use the mean value of .82, which yields a correction factor of 1.10.

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 (1998a, 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 population. The standard deviations can be compared by dividing the study population standard deviation by the reference group population standard deviation, that is $u = SD_{study} / SD_{ref}$. 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; the average standard deviation of *g* loadings is .128. The nine NIT subtests' *g* loadings derived from Must et al. (2003b) have an $SD = .0943$, which gives a $u = .737$, yielding a correction factor of 1.36. The values are slightly different in the case of *g* loadings derived using the alternative CFA method employed in Must et al. (2009). For the 1930s' cohort the $SD = .134$ is optimally comparable to the reference study *SD*, hence no correction is made to this cohort. For the 2006 cohort the $SD = .115$ is slightly lower than the optimal value, hence a small correction of $u = .898$, or 1.11. As already noted by Jensen (1998a) restriction of range strongly attenuates vector correlations.

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 resulting in a highly *g*-saturated composite score. Jensen (1998a, 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 a somewhat lower *g* saturation. Jensen (1998a. 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 likely that most empirical *g* loadings will underestimate "true" *g* loadings; so, empirical *g* loadings correlate about .85 with "true" *g* loadings. To limit the risk of overcorrection, we conservatively chose the value of .90 for the correction, yielding a correction factor of 1.11.

Table 1 shows the four statistical corrections and their values. Combining the four corrections leads to an overall correction factor of $1.05 \times 1.10 \times 1.36 \times 1.11 = 1.74$. We also report the correction factor (1.11) for the restriction of range in *g* loadings amongst the subtests comprising the CFA derived estimates in the 2006 cohort reported in Must et al. (2009). This means that there is a quite substantial amount of statistical error in the correlations of *g* loadings with the various gains. Carrying out the corrections leads to a less obstructed view of the relationships at the construct level.

Table 1
Psychometric corrections applied to the correlation between *g* loadings and the secular gains corrected for the secular gains in guessing.

Corrections	Correction factor
Reliability of *g* vector	1.05
Reliability of gains vector	1.10
Restriction of range in *g* loadings	1.36/1.11
Deviation from perfect construct validity	1.11

3. Results

Table 2 presents the CFA derived *g* loadings for both cohorts, along with the numbers of wrong answers and a measure of the percentage impact on subtest scores of false positive answers.

Table 3 reveals that there are significant and large magnitude correlations between the *g* loadings of subtests and the numbers of wrong answers in the case of both cohorts, indicating a robust association between subtest complexity and difficulty. This relationship is especially strong in the 2006 cohort (.73 — becomes .94 with correction), and is significantly so ($z = -21.47$), which suggests that *g* is more determinative of subtest difficulty in this cohort. The *g* loading positively correlates with the % impact of false positive answers in both cohorts, however the difference between the *r*'s is also significant favoring the latter-born cohort ($z = -19.27$), suggesting that *g* becomes more sensitive to the impact of false positive answers in the latter born cohort. An interesting observation concerned the relationship between the % impact of false positives and the number of wrong answers. The relationship switches sign from negative to positive across cohorts, perhaps reflecting the possibility that in latter born cohorts, guessing may be both strategic and more frequent (as evidenced by the decline in missing answers between cohorts [Must & Must, 2013]), hence it leads to an increase in both right and wrong answers. Under conditions of guessing in which more items are attempted we would expect that the error rate becomes more strongly coupled to subtest *g* loadings, which in turn correspond more strongly to overall subtest difficulty.

On the basis of the pattern of correlations we are justified in our assumption that secular gains in correct answers due to guessing (i.e. the Brand effect) should generally occur on subtests comprised of items that are simultaneously more cognitively complex (i.e. more *g* loaded) and more difficult (i.e. elicit a higher proportion of wrong answers). This will be explicitly tested in the subsequent analysis.

Table 4 presents all data on *g* loadings and the various secular gains used in this analysis. The results indicate that the secular gains show an anti-Jensen effect of modest magnitude; this value of the observed $r = -.18$ is virtually identical to the sample-size weighted observed $r = -.17$ from the large psychometric meta-analysis of te Nijenhuis and van der Flier. As expected, the correlation between *g* loadings and the secular gains corrected for guessing are much stronger, namely an observed $r = -.47$. Correcting this observed *r* for four statistical artifacts yields a $Rho = -.82$.

The correlation between *g* loadings and the Brand effect is effectively positively monotonic at $r = .95$. As predicted, harder items are more likely to evoke guessing, and are therefore more sensitive to the Brand effect. The Brand effect is an almost perfect Jensen effect.

4. Discussion

The causes of secular IQ gains is one of the biggest puzzles in intelligence research. Brand's (1996) suggestion that increased guessing is a potentially important cause and this was supported in recent studies by Must and Must (2013) and Pietschnig et al. (2013). The meta-analysis by te Nijenhuis and van der Flier (2013) shows that there is a quite modest negative correlation between *g* loadings and the magnitude of secular gains. We hypothesized that correcting for increased guessing would make this correlation substantially more negative, which would mean that, once controlled for the Brand effect, the authentic Flynn effect is even more independent of *g* than previously thought.

Consistent with the results of psychometric meta-analysis (te Nijenhuis & van der Flier, 2013), the pattern of secular gains is here modestly associated with an anti-Jensen effect. Correcting the anti-Jensen effect on the secular gains for the increase in guessing leads to a stronger negative association, as predicted. Correcting for four sources of measurement error makes the association very strong: it is actually quite close to the value of −1. This indicates that there are indeed two clear sources of secular gains on IQ tests. On the one hand there is a strongly environmentally-driven authentic Flynn effect which is very respectively negatively related to the *g*-loadedness of subtests. As the *g* loadings of subtests are

Table 2
The CFA derived *g* loadings for both the 1933–36 ($N = 890$) and 2006 ($N = 913$) cohorts along with the number of wrong answers and also the percentage difference between the number of right answers and right answers adjusted for guessing — measuring the % impact of false positives on subtest scores.

Subtest	CFA 1933–36 *g* loadings	Wrong answers	% false positive	CFA 2006 *g* loadings	Wrong answers	% false positive
Arithmetical reasoning (A1)	.69	2.5	45.05	.62	3.2	50
Sentence completion (A2)	.82	2.3	22.44	.74	2.9	20.94
Same–different (A4)	.74	2.9	10.6	.66	3.4	10.06
Symbol–digit (A5)	.52	.71	.96	.45	.51	.49
Computation (B1)	.51	3	29.41	.57	3	31.92
Information (B2)	.88	5.8	30.05	.74	10.8	64.67
Vocabulary (B3)	.67	3.8	14.3	.62	5.5	18.87
Analogies (B4)	.74	8	84.21	.78	8.1	50.78
Comparisons (B5)	.53	2.1	7.82	.48	2.4	6.43

Table 3
Correlation matrix amongst *g* loadings, the number of wrong answers and the % impact of false positive answers. The 1933–36 cohorts are below the diagonal and the 2006 cohorts are above. For all vector correlations involving subtest *g* loadings meta-analytic corrections for reliability, range restriction amongst the *g* loadings and the validity coefficient are made. Both the raw and corrected correlations (in parentheses) are reported.

	CFA *g* loadings (1933–36)	Wrong answers	% false positive impact
CFA *g* loadings (2006)	.89 (1)	.73 (.94)	.67 (.87)
Wrong answers	.53 (.62)	1	.79
% false positive impact	.35 (.40)	−.53	1

All coefficients are statistically significant at $P < .05$.

very strongly positively related to their heritability (Rushton & Jensen, 2010), this strengthens the argument for the strong environmentality of the Flynn effect. On the other hand there is a clear Brand effect, which results from secular increases in the degree to which guessing is being favored as a strategy for dealing with subtests containing more difficult items.

This is a potentially important finding as Pietschnig et al. (2013) note that Brand's hypothesis has seldom been tested and therefore represents a large potential 'unknown' in studies of the Flynn effect. Furthermore its theoretical implications for the relationship between secular gains and *g* were never addressed in the original exchange of papers between Brand and Flynn (e.g. Brand, 1990; Flynn, 1990) leaving interesting questions unanswered. As we have already observed, our finding also has potentially significant implications for causal theories of secular gains. Flynn (2009) for example has argued that these gains do not represent actual increases in general intelligence, but are instead associated with the proliferation of heuristics and habits of thought, which adapt populations to the cognitive demands of modernity. This theory necessitates that secular gains be almost completely independent of *g*, which is essentially what we find when the Brand effect is controlled. Similarly the co-occurrence model of dysgenics and the Flynn effect argues that dysgenic fertility is concentrated on *g*, hence highly heritable and culturally neutral measures of cognitive ability should in fact be trending negatively over time consistent with selection pressures, whereas less heritable and more culturally and environmentally sensitive abilities should be trending upwards, tracking the improving environment (Woodley, 2012; Woodley & Meisenberg, 2013; Woodley et al., 2013). This model also requires that there should be no secular gains at the level of *g*, which cannot be simultaneously rising and falling. The present finding is therefore strongly in line with the expectations of the co-occurrence model.

The almost perfect Jensen effect on the Brand effect is theoretically interesting for two reasons. Firstly its existence falsifies the strict biological vs. cultural/environmental continuum interpretation of the Jensen/anti-Jensen effect nexus (Rushton, 1999; te Nijenhuis, 2013). Whilst the relationships between most variables and *g* loadings align consistently with predictions from this model, the Brand effect is a clear exception — being both purely cultural in origin and hollow (in the correct sense — i.e. having no criterion validity) whilst appearing to have its main effect on *g*. This in turn makes the idea that small cross-loadings of secular gains on genetic factors in principal components analysis (Rushton, 1999) may indicate modest contributions from factors such as heterosis less plausible (e.g. te Nijenhuis & van der Flier, 2013), as the pattern of cross-loadings could just as easily stem from uncontrolled Brand effects.

Moreover the existence of the Brand effect has implications for the dysgenic hypothesis as the co-occurrence model argues that dysgenic effects are concentrated on heritable *g*. Hence if the Brand effect is inflating secular gains by giving them a pseudo-*g* component, then the presence of this effect across the decades might have made psychometric tests even more insensitive to dysgenic effects than was previously thought. This strongly reinforces the rationale for using temporal trends in culturally neutral and highly biological measures of general intelligence such as reaction times as a means of evaluating the real impacts of dysgenic fertility on population intelligence (Woodley et al., 2013). It needs to be noted however, that this interpretation is being challenged (e.g. Dodonova & Dodonova, 2013; Flynn, 2013; Nettelbeck, 2014; Silverman, 2013).

Finally, whilst the results hint at the total independence of the authentic Flynn effect from *g*, they do not directly demonstrate this. The *Rho* value was −.82, which leaves a small residual variance potentially associated with *g* in our study

Table 4
Data on the NIT's *g* loadings (from the 1930s' cohort in Must et al., 2003b), secular gains between 1933 and 2006, secular gain corrected for guessing, and Brand effect or gain due to guessing; the last three effects are expressed in standard deviations (*d*).

Subtest	*g* loading	Gain	Gain corrected for guessing	Gain due to guessing
Arithmetical reasoning (A1)	.782	.545	.152	.393
Sentence completion (A2)	.845	1.238	.759	.479
Same–Different (A4)	.669	1.072	.815	.275
Symbol–Digit (A5)	.669	1.654	1.621	.033
Computation (B1)	.623	−.333	−.225	−.108
Information (B2)	.885	−.016	−.667	.651
Vocabulary (B3)	.780	.739	.301	.438
Analogies (B3)	.809	1.092	.631	.461
Comparisons (B5)	.653	1.710	1.609	.101
Correlation with *g*		−.180*	−.471*	.945*

* $p < .05$.

(33%). If the co-occurrence model ultimately proves to be wrong, or if this variance cannot be accounted for by some other currently unknown effect, then this suggests that *g* might yet play a role in the Flynn effect, perhaps in response to social multiplication factors that specifically work on the environmentality of *g* (Colom et al., 2001), or via improved nutrition (Lynn, 1989). Maybe the small disparity could even be accounted for by heterosis (Mingroni, 2007).

Future research should aim to replicate the findings of Must and Must (2013) and also the present paper in a wider array of test batteries. This will give a better idea of the aggregate effect of the correction that needs to be made to the results of future meta-analyses on the relationship between secular IQ gains and test *g* loadings.

References

Arend, I., Colom, R., Botella, J., Contreras, M.Ē. J., Rubio, V., & Santacreu, J. (2003). Quantifying cognitive complexity: evidence from a reasoning task. *Personality and Individual Differences*, *35*(3), 659–669.

Armstrong, E. L., & Woodley, M. A. (2014). The rule-dependence model explains the commonalities between the Flynn effect and IQ gains via retesting. *Learning and Individual Differences*, *29*, 41–49.

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

Brand, C. R. (1987a). Intelligence testing: Bryter still and Bryter? *Nature*, *328*, 110.

Brand, C. R. (1987b). British IQ: Keeping up with the times. *Nature*, *328*, 761.

Brand, C. R. (1990). A 'gross' underestimate of a 'massive' IQ rise? A rejoinder to Flynn. *Irish Journal of Psychology*, *11*, 52–56.

Brand, C. R. (1996). *The g factor: General intelligence and its implications.* Chichester: Wiley.

Brand, C. R., Freshwater, S., & Dockrell, N. (1987). Has there been a 'massive' rise in IQ levels in the West? Evidence from Scottish children. *Irish Journal of Psychology*, *10*, 388–393.

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.

Coyle, T. R., & Pillow, D. R. (2008). SAT and ACT predict college GPA after removing *g*. *Intelligence*, *36*, 719–729.

Dickens, W. T., & Flynn, J. R. (2001). Heritability estimates versus large environmental effects: The IQ paradox resolved. *Psychological Review*, *108*, 346–369.

Dodonova, Y. A., & Dodonova, Y. S. (2013). Is there any evidence of historical slowing of reaction time? No, unless we compare apples and oranges. *Intelligence*, *41*, 674–687.

Dolan, C. V. (2000). Investigating Spearman's hypothesis by means of multi-group confirmatory factor analysis. *Multivariate Behavioral Research*, *35*, 21–50.

Eppig, C., Fincher, C. L., & Thornhill, R. (2010). Parasite prevalence and the worldwide distribution of cognitive ability. *Proceedings of the Royal Society Series B: Biological Sciences*, *277*, 3801–3808.

Flynn, J. R. (1984). The mean IQ of Americans: Massive gains 1932 to 1978. *Psychological Bulletin*, *95*, 29–51.

Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin*, *101*, 171–191.

Flynn, J. R. (1990). Massive IQ gains on the Scottish WISC: Evidence against Brand et al.'s hypothesis. *The Irish Journal of Psychology*, *11*, 41–51.

Flynn, J. R. (2009). *What is intelligence?* (2nd ed.). New York: Cambridge University Press.

Flynn, J. R. (2012). *Are we getting smarter? Rising IQ in the 21st century.* New York: Cambridge University Press.

Flynn, J. R. (2013). The Flynn effect and "Flynn's paradox". *Intelligence*, *41*, 851–857.

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

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

Herrnstein, R. J., & Murray, C. (1994). *The bell curve: Intelligence and class structure in American life.* New York: Free Press.

Howard, R. W. (2001). Searching the real world for signs of rising intelligence. *Personality and Individual Differences*, *30*, 1039–1058.

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.

Husén, T., & Tuijnman, A. (1991). The contribution of formal schooling to the increase in intellectual capital. *Educational Researcher*, *20*, 17–25.

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

Jensen, A. R. (1998b). Adoption data and two *g*-related hypotheses. *Intelligence*, *25*, 1–6.

Lynn, R. (1989). A nutrition theory of the secular increases in intelligence — Positive correlations between height, head size and IQ. *The British Journal of Educational Psychology*, *59*, 372–377.

Mingroni, M. A. (2004). The secular rise in IQ: Giving heterosis a closer look. *Intelligence*, *32*, 65–83.

Mingroni, M. A. (2007). Resolving the IQ paradox: Heterosis as a cause of the Flynn effect and other trends. *Psychological Review*, *114*, 829–896.

Must, O., & Must, A. (2013). Changes in test-taking patterns over time. *Intelligence*, *41*, 791–801.

Must, O., Must, A., & Raudik, V. (2003a). The Flynn effect for gains in literacy found in Estonia is not a Jensen effect. *Personality and Individual Differences*, *34*, 1287–1292.

Must, O., Must, A., & Raudik, V. (2003b). The secular rise in IQs: In Estonia the Flynn effect is not a Jensen effect. *Intelligence*, *31*, 461–471.

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

Nagoshi, G. T., & Johnson, R. C. (1986). The ubiquity of *g*. *Personality and Individual Differences*, *7*, 201–207.

Nettelbeck, T. J. (2014). Smarter but slower? A comment on Woodley, te Nijenhuis and Murphy (2013). *Intelligence*, *42*, 1–4.

Nevin, R. (2000). How lead exposure relates to temporal changes in IQ, violent crime, and unwed pregnancy. *Environmental Research*, *83*, 1–22.

Pietschnig, J., Tran, U. S., & Voracek, M. (2013). Item-response theory modeling of IQ gains (the Flynn effect) on crystallized intelligence: Rodgers' hypothesis yes. Brand's hypothesis perhaps. *Intelligence*, *41*, 791–801.

Reeve, C. L., Lyerly, J. E., & Peach, H. (2013). Adolescent intelligence and socioeconomic wealth independently predict adult marital and reproductive behavior. *Intelligence*, *41*, 358–365.

Rushton, J. P. (1999). Secular gains 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., & Jensen, A. R. (2010). 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.

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

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. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology*, *62*, 529–540.

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

Silverman, I. W. (2010). Simple reaction time: It is not what it used to be. *American Journal of Psychology*, *123*, 39–50.

Silverman, I. W. (2013). Testing the hypothesized effect of dysgenic fertility on intelligence with existing reaction time data: A comment on Woodley, te Nijenhuis, and Murphy (2013). *Intelligence*, *41*, 664–666.

te Nijenhuis, J. (2013). The Flynn effect, group differences and *g* loadings. *Personality and Individual Differences*, *55*, 224–228.

te Nijenhuis, J., & van der Flier, H. (2013). Is the Flynn effect on *g*?: A meta-analysis. *Intelligence*, *41*, 802–807.

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.

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

Woodley, M. A. (2012). The social and scientific temporal correlates of genotypic intelligence and the Flynn effect. *Intelligence*, *40*, 189–204.

Woodley, M. A., & Meisenberg, G. (2013). A Jensen effect on dysgenic fertility: An analysis involving the National Longitudinal Survey of Youth. *Personality and Individual Differences*, *55*, 279–282.

Woodley, M. A., te Nijenhuis, J., & Murphy, R. (2013). Were the Victorians cleverer than us? The decline in general intelligence estimated from a meta-analysis of the slowing of simple reaction time. *Intelligence*, *41*, 843–850.