Intelligence 46 (2014) 209–215

Contents lists available at ScienceDirect

Intelligence

Are Headstart gains on the *g* factor? A meta-analysis

Jan te Nijenhuis [a,*], Birthe Jongeneel-Grimen [b], Emil O.W. Kirkegaard [c]

[a] University of Amsterdam, Work and Organizational Psychology, The Netherlands
[b] University of Amsterdam, Amsterdam Medical Center, The Netherlands
[c] University of Århus, Department of Linguistics, Denmark

ARTICLE INFO

Article history:
Received 12 March 2014
Received in revised form 21 May 2014
Accepted 1 July 2014
Available online xxxx

Keywords:
Headstart
Intelligence
Jensen effect
Meta-analysis
Compensatory education

ABSTRACT

Headstart studies of compensatory education tend to show impressive gains on IQ scores for children from low-quality environments. However, are these gains on the *g* factor of intelligence? We report a meta-analysis of the correlation between Headstart gains on the subtests of IQ batteries and the *g* loadings of these same subtests ($K = 8$ studies, total $N = 602$). A meta-analytic sample-weighed correlation of −.51 was found, which became −.80 after corrections for measurement error. We conclude that the pattern in Headstart gains on subtests of an IQ battery is highly similar to the pattern in test–retest gains and is hollow with respect to *g*. So, Headstart leads to gains in IQ scores, but not to gains in *g*. We discuss this finding in relation to the Flynn effect, training effects, and heritability.

1. Introduction

Changes in IQ scores are one of the big puzzles of intelligence research. Minor changes are often due to measurement error, but this is unlikely to be the cause of more substantial fluctuations. For instance, raw scores on standard IQ tests have been going up for decades (the Flynn effect; Lynn, 2013), and as the effect is large and unidirectional, simple measurement error does not offer sufficient explanatory power. Much research in the past three decades has been centered on the Flynn effect, e.g. the recent special issue in *Intelligence* (Thompson, 2013). The nature of the effect is hotly debated. Some authors, like Lynn (2013), believe it to be a real increase in intelligence, citing, among other things, the similar rise in height as evidence. Many non-specialists similarly treat the Flynn effect as a real increase in intelligence (e.g. Somin, 2013). Hypothesized causes for a real increase include: better nutrition (Flynn, 1987; Lynn, 2006), heterosis (i.e. outbreeding, Mingroni, 2007), improvement in hygiene (Eppig, Fincher, & Thornhill, 2010), and reduced lead poisoning (Nevin, 2000).

An alternate explanation posits that the effect has little or nothing to do with general intelligence, or *g*, itself. Jensen (1998, p. 143) invented the method of correlated vectors to check whether a phenomenon has to do with the underlying latent variable of interest, i.e. *g*, or whether it has to do with the non-*g* variance. Other researchers have since called phenomena that show a positive relation to the *g* loading of subtests "Jensen effects" (e.g. Colom, Juan-Espinosa, & García, 2001; Rushton, 1998). Wholly or partly genetically influenced variables, such as subtest heritabilities (Rushton & Jensen, 2010), dysgenic fertility (Woodley & Meisenberg, 2013), fluctuating asymmetry (Prokosch, Yeo, & Miller, 2005), brain size (Rushton & Ankney, 2009), inbreeding depression (Jensen, 1998), and reaction times (Jensen, 1998) have been shown to be Jensen effects.

On the other hand, environmental variables seem to be negative Jensen effects. te Nijenhuis and van der Flier (2013) reported a meta-analysis of the Flynn effect which yielded a negative Jensen effect of −.38 (after corrections). Moreover, in a newer study, Woodley, te Nijenhuis, Must, and Must

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

http://dx.doi.org/10.1016/j.intell.2014.07.001
0160-2896/

(2014) reexamined one of the datasets in this meta-analysis and found that if one corrects for increased guessing at the harder items (the Brand effect) then the negative Jensen effect came even closer to −1 at −.82, indicating that the gains may be more hollow with respect to *g* than previously thought (see also Flynn, te Nijenhuis, & Metzen, 2014).

In a related study, te Nijenhuis, van Vianen, and van der Flier (2007) reported a meta-analysis of 64 studies (total $N = 26{,}990$) on score gains from test training yielding a negative Jensen effect of −1.0 (after corrections). Score gains from training are theoretically interesting because they present a clear case that one can increase the proxy (or manifest variable), IQ, without increasing the underlying latent variable of interest, *g*. Whatever causes the Flynn effect gains, it seems likely this effect is similarly mostly hollow with respect to *g*; it represents no large gain in *g*. Accordingly, we have not seen the substantial increase in the number of geniuses in Western countries that we could expect to result from a mean increase in *g* of a standard deviation or more (Jensen, 1987, pp. 445–446). As Herrnstein and Murray (1994, p. 364) point out, a mere 3 IQ point increase in *g* would make a large difference on the tails of the distribution. For instance, it would increase the number of people above IQ = 130, often taken as the threshold of giftedness, by 68% (from 2.3% to 3.6%). An increase of one or more *SD* in *g* could not possibly be overlooked.

1.1. Compensatory education and IQ gains: g-loaded?

The largest program for compensatory education is Project Headstart, which began as a program to improve intellectual functioning and to increase academic achievement (Caruso, Taylor, & Detterman, 1982) and has been running since 1965. It is a public preschool program that was designed for disadvantaged children to close the achievement gaps between the disadvantaged child and their more advantaged peers (Soriano, Duenas, & LeBlanc, 2006). The program is massive, involving 1 million children, and cost almost 8 billion dollars in 2012 (U.S. Department of Health & Human Services, 2012).

Several meta-analyses of Headstart studies showed that children in the program outscored children in control groups (Caruso et al., 1982; Ramey, Bryant, & Suarez, 1985; Nelson, Westhues, & MacLeod, 2003; see also Protzko, Aronson, & Blair, 2013). However, no one, to our knowledge, has yet carried out an analysis to see if the gains are a Jensen effect.

In 1969, when Jensen published his famous article "How much can we boost IQ and scholastic achievement?" (Jensen, 1969) he drew the conclusion that compensatory education had been tried and had failed. Although initial IQ gains were sometimes large, they diminished with time and so could not be expected to close the gaps between racial and economic groups. Spitz (1986) reviewed most of the literature on the attempts to increase intelligence and his conclusions were also mostly negative. He mentions (p. 103) that in the Perry Preschool Program, the teachers seemed to focus on teaching material that was similar to the content of subtests of the IQ tests, so-called "teaching to the test". It is not unlikely that highly comparable practices were present in many other programs, including Headstart.

In the widely accepted model in Fig. 1 U_n is the variance specific to each subtest, V_n. The teaching to the test-hypothesis can be clearly stated in terms of the model. According to the hypothesis, when one trains test takers on the exact subtests or subtests very similar to those used in a test, the resultant effect is on the U_n factors in the model (and maybe somewhat on the group factors F_n), but there is no increase in the latent variable *g*. If one assumes that test takers are taught comparably on all the subtests, then this leads directly to the prediction that any resultant training effect should have a strong negative correlation with the *g* loading of the subtests. This is because, for each V_n, the greater the influence of U_n, the smaller the influence of *g* (through the group factors). If ability in U_n is increased, it will be higher on the V_ns where *g* has a smaller influence, that is, that are less *g*-loaded (see also Jensen, 1998, pp. 336–337).

This leads us to the present study. The goal was to determine whether the gains from Headstart are similar to training effects, with a strong negative Jensen effect, or whether they are genuine increases in *g*, in which case they should show a strong Jensen effect.

2. Method

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. The results of perfectly conducted studies would allow a less obstructed view of the underlying construct–level relationships (Schmidt & Hunter, 1999). The goal of the present psychometric meta-analysis is to provide reliable estimates of the true correlation between Headstart gains and the magnitude of *g* loadings. As the techniques we use are relatively unknown to the majority of readers we choose to give a detailed description of the techniques. However, highly similar descriptions have also been used in other recent publications.

2.1. Searching and screening studies

To identify studies for inclusion in the meta-analysis, both electronic and manual searches for studies that contained cognitive ability data of Headstart children or adults who participated in a Headstart program as a child were conducted in 2007. Four methods were used to obtain Headstart gains from both published and unpublished studies for the present meta-analysis. First, an electronic search of published research was conducted, using PsycINFO, ERIC, PiCarta, Academic Search Premier, Web of Science, and PubMed. The following keyword combinations were used to conduct searches: Headstart, Head Start, preschool children, and kindergarten children in combination with the keywords IQ, intelligence, intellectual development, *g*, GMA, general mental ability, cognitive development, cognitive ability, and general cognitive ability. Second, we browsed the tables of contents of several major research journals of education, development, and of intelligence, such as *American Educational Research Journal* 1968–2007, *Journal of Educational Research* 1965–2007, *Intelligence* 1977–2007, *Psychological Science* 1990–2007, *Child Development* 1930–2007, and *Developmental Psychology* 1969–2007. Third, several well-known researchers who have

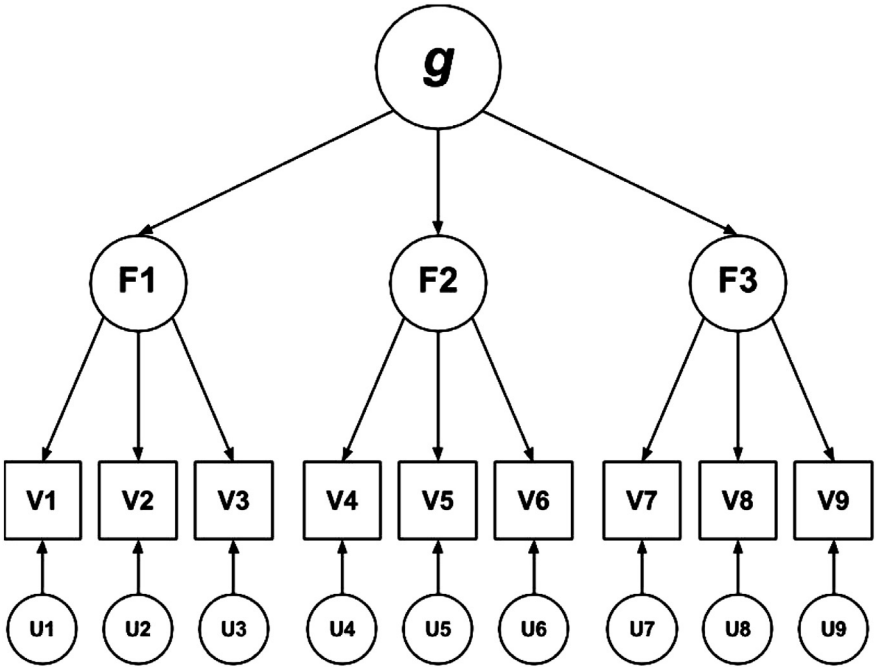

Fig. 1. Hierarchical model of human cognitive abilities, using a simplified form of the model in Jensen and Weng (1994). Circles are latent factors and squares are manifest variables.

conducted cognitive ability research of Headstart, preschool, and kindergarten children or adults who participated in a Headstart program, preschool, or kindergarten as a child were contacted in order to obtain any additional articles or supplementary information. Finally, we checked the reference list of all currently included empirical studies to identify any potential articles that may have been missed by earlier search methods.

2.2. Inclusion rules

Studies that reported IQ scores of Headstart children, preschool, and kindergarten children were included in the meta-analysis. We used the term "Headstart" in a generic sense, so it included preschool and kindergarten children as well. For a study to be included in the meta-analysis two criteria had to be met: First, to get a reliable estimate of the true correlation between Headstart gains and the *g* loadings the cognitive batteries had to have a minimum of seven subtests; second, well-validated tests had to be used. The general inclusion rules were applied and yielded six papers which resulted in eight correlations between *g* and *d* (Headstart gains).

2.3. Computation of Headstart gains

One of the goals of the present meta-analysis is to obtain a reliable estimate of the true correlation between Headstart gains (*d*) and *g*. To be able to compute *d* (Headstart gains) we needed to compare the results of the intervention groups against the results of comparison groups. A limitation of all the studies included in this meta-analysis is that none of them included a comparison group. In general, FACES (Head Start Family and Child Experiences Survey study) students entered the program with measures of vocabulary, letter recognition, and math that were about one-half to a full standard deviation below the national average (see Zill, Resnick, Kim, O'Donnell, & Sorongon, 2003). We therefore decided to compare the mean of the scaled scores of Headstart children with an artificially generated comparison group with total IQ scores one *SD* below the mean of the scaled scores of the standardization groups of the particular test in question, because such a simulated comparison group is cognitively more similar to the Headstart children than the national standardization groups. Headstart, preschool, and kindergarten gains (*d*) were computed by subtracting the mean of the comparison group from the mean of the intervention group. The result was then divided by the (mean) *SD* of the standardization group(s) of the particular test in question.

2.4. Computation of g loadings

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). Finally, Pearson correlations between Headstart gains and the *g* loadings were computed.

2.5. Corrections for artifacts

Psychometric meta-analytical techniques (Hunter & Schmidt, 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 that alter the value of outcome measures listed by Hunter and Schmidt (2004). These are: (1) sampling error, (2) reliability of the vector of *g* loadings, (3) reliability of the vector of Headstart gains (*d*), (4) restriction of range of *g* loadings, and (5) deviation from perfect construct validity.

2.5.1. 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.

2.5.2. Correction for reliability of the vector of g loadings

The values of *r* (*g* × Headstart gains) are attenuated by the reliability of the vector of *g* loadings for a given battery. When two samples have a comparable *N*, the average correlation between vectors is an estimate of the reliability of each vector. Several samples were compared that differed little on background variables. For the comparisons using children we chose samples that were highly comparable with regard to age. Samples of children in the age of 3 to 5 years were compared against other samples of children who did not differ more than 0.5 years of age. Samples of children in the age of 6 to 17 years were compared against other samples of children who did not differ more than 1.5 years of age. For the comparisons of adults we compared samples in the age of 18 to 95 years.

We collected correlation matrices from test manuals, books, articles, and technical reports. The large majority came from North America, with a large number of European countries, and also a substantial number from Korea, China, Hong Kong, and Australia. This resulted in about 700 data points, which led to 385 comparisons of *g* loadings of comparable groups which provided an indication of the reliability for that group. To give an illustration of the procedure, van Haasen et al. (1986) report correlation matrices of the Dutch and the Flemish WISC-R for 22 samples in the age of 6–16 years. We compared samples of children in the age of 6 to 17 years with other samples of children who do not differ by more than 1.5 years. Because the samples of children reported in van Haasen et al. (1986) were between 6 and 17 years we only compared children who did not differ more than 1.5 years. The *N*s in these samples were comparable. This resulted in an average correlation of .78 (combined $N = 3018$; average $N = 137$).

A scatter plot of reliabilities against *N*s should show that the larger *N* becomes, the higher the value of the reliability coefficients, with an asymptotic function between *r* (*g* × *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.

2.5.3. Correction for reliability of the vector of Headstart gains (d)

The values of *r* (*g* × Headstart gains) are attenuated by the reliability of the vector of Headstart 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 Headstart gains was estimated using the present datasets and by comparing the samples that took the same test and that were comparable with regard to age and sample size. As an illustration of the procedure, consider the vectors of Headstart gains from datasets on the WPPSI. McNamara, Porterfield, and Miller (1969) tested children ($N = 42$) with an average age of 5.8 years (age range 4.8 to 6.6 years); Yater, Barclay, and Leskosky (1971) tested children ($N = 48$) with an average age of 5.3 years (age range 4.8 to 6.0 years); and Henderson and Rankin (1973) tested children ($N = 49$) with an estimated mean age of 5.5 years (age range 5.0 to 6.0 years). The correlations between the *d* vectors of the three studies are respectively .90 (total $N = 90$; average $N = 45$), .64 (total $N = 97$; average $N = 49$), and .72 (total $N = 91$; average $N = 46$). Lowe, Anderson, Williams, and Currie (1987) also tested children ($N = 169$) on the WPPSI. They had an average age of 5.9 years (age range 5.6 to 6.2 years). We decided not to compare vectors of Headstart gains from the dataset in Lowe et al. (1987) because the differences in sample size were too large.

An asymptotic function between *r* (*d* × *d*) and *N* is expected. We checked to see which curve gave the best fit to the expected asymptotic function. Fig. 2 presents the scatter plot of the reliability of the vector of Headstart gains and sample size, and the curve that fitted optimally.

2.5.4. Correction for restriction of range of g loadings

The values of *r* (*g* × Headstart gains) are attenuated by the restriction of range of *g* loadings in many of the standard test batteries. The most highly *g*-loaded batteries tend to have the smallest range of variation in the subtests' *g* loadings. Jensen (1998, pp. 381–382) showed that restriction in the magnitude of *g* loadings strongly attenuates the correlation between *g* loadings and standardized group differences. Hunter and Schmidt (1990, pp. 47–49) state that the solution to 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 standard deviation of the study population by the standard deviation of the reference group, that is $u = SD_{study} / SD_{ref}$. As references we used tests that are broadly regarded as exemplary for the measurement of the intelligence domain, namely the various versions of the Wechsler tests for children and adults. The average standard deviation of *g* loadings of the various versions of the Wechsler Bellevue (W-B), Wechsler Preschool and Primary Scale of Intelligence (WPPSI), Wechsler Intelligence Scale for Children (WISC), Wechsler Intelligence Scale for Children — Revised (WISC-R), Wechsler Intelligence Scale for Children — Third Edition (WISC-III), and the Wechsler Intelligence Scale for Children — Fourth Edition (WISC-IV) from datasets from countries all over the world was 0.132. We used this value as our reference in the studies with children. The average standard deviation of *g* loadings of the various versions of the Wechsler Adult Intelligence Scale (WAIS), Wechsler Adult Intelligence Scale — Revised (WAIS-R), and the

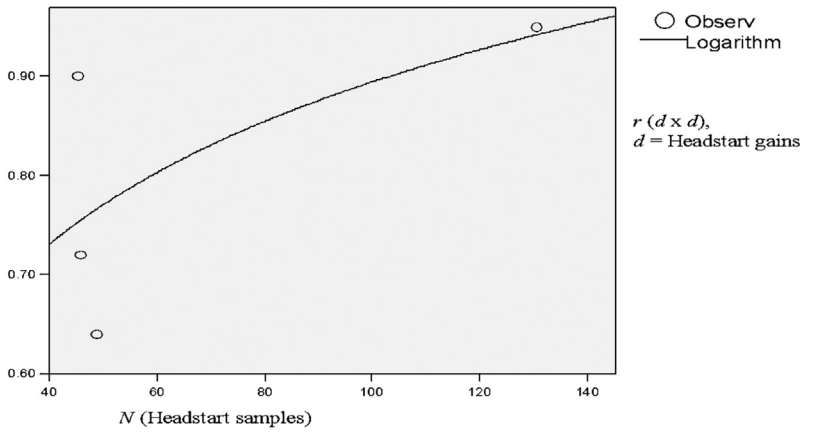

Fig. 2. Scatter plot of reliability of the vector of Headstart gains and sample size and regression line.

Wechsler Adult Intelligence Scale — Third Edition (WAIS-III) from datasets from countries all over the world was 0.107. This was used as the reference value in the studies with adults. In so doing, the *SD* of *g* loadings of all test batteries was compared to the average *SD* in *g* loadings in the Wechsler tests for, respectively, children and adults.

2.5.5. Correction for deviation from perfect construct validity

The deviation from perfect construct validity in *g* attenuates the values of *r* (*g* × Headstart gains). 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* (Johnson, Bouchard, Krueger, McGue, & Gottesman, 2004; Johnson, te Nijenhuis, & Bouchard, 2008). The values of *r* (*g* × Headstart gains) are 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 is of the composite score. The Wechsler tests have a large number of subtests with quite high *g* loadings. This yields 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 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 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 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 was carried out on the values of *r* (*g* × Headstart gains) after correction for the first four artifacts. To limit the risk of over-correction, we conservatively chose the value of .90 for the correction.

3. Results

The results of the studies on the correlation between *g* loadings and Headstart gains are shown in Table 1. The table gives data derived from six studies, with participants numbering a total of 602. It presents the reference for the study, the cognitive ability test used, the correlation between *g* loadings and Headstart gains, the sample size, and the mean age (and range of age). It is clear that all these correlations are negative and about half quite strongly.

Table 2 lists the results of the psychometric meta-analysis of the eight data points. The estimated true correlation has a value of −.72, and artifacts explain 71% of the variance in the observed correlations. Finally, a correction for deviation from perfect construct validity in *g* was made, using a conservative value of .90. This resulted in a value of −.80 for the final estimated true negative Jensen effect.

Table 1
Studies of correlations between *g* loadings and Headstart gains.

Reference	Test	*r*	*N*	Age mean (range)
McNamara et al. (1969)	WPPSI	−.391	42	5.80 (4.80–6.60)
Yater et al. (1971)	WPPSI	−.298	48	5.30 (4.80–6.00)
Henderson and Rankin (1973)	WPPSI	−.284	49	5.50[a] (5.00–6.00)
Lowe et al. (1987)	WPPSI	−.386	169	5.90 (5.60–6.20)
Lowe et al. (1987)	WISC-R	−.770	94	9.80 (9.50–10.20)
Lowe et al. (1987)	WAIS-R	−.356	40	17.40 (17.10–17.80)
Krohn, Lamp, and Phelps (1988)	K-ABC[b]	−.757	38	4.25 (3.30–4.75)
Gridley, Miller, Barke, Fischer, and Smith (1990)	K-ABC[b]	−.665	122	4.50 (3.17–5.42)

Note. In general, the *g* loadings were based on the correlation matrix taken from test manuals or from the correlation matrix based on the largest sample size we could find. A detailed description of all the data points for this meta-analysis can be found in the Supplementary material.

[a] Estimated mean age.

[b] Kaufman Assessment Battery for Children.

Table 2
Meta-analytical results for correlation between Headstart gains and *g* loadings after corrections for reliability and restriction of range.

K	*N*	*r*	*SDr*	Rho-4	*SD*rho-4	Rho-5	%VE	80% CI
8	602	−.51	.16	−.72	.01	−.80	71	−.58 to −.85

Note. *K* = number of correlations; *N* = total sample size; *r* = mean observed correlation (sample size weighted); *SDr* = standard deviation of observed correlation; rho-4 = observed correlation (corrected for unreliability and range restriction); *SD*rho = standard deviation of true correlation; rho-5 = observed correlation (corrected for unreliability, range restriction, and imperfect measurement of the construct);%VE = percentage of variance accounted for by artifactual errors; 80% CI = 80% credibility interval.

4. Discussion

Studies of compensatory education sometimes show impressive gains on IQ scores for children from low-quality environments. Are the Headstart gains similar to training effects and the Flynn effect, showing a strong negative Jensen effect, or are they genuine increases in *g*, in which case they should show a strong Jensen effect?

Results were strongly in line with the prediction that Headstart involves a lot of teaching to the test, so that the gains would be strongly at the level of the specific or group factors. The gains involve mostly the non-*g* variance, which means that they were mostly hollow in terms of *g*. The final estimated true correlation of −.80, rather than a correlation of exactly −1.0, need not mean that there was some gain in *g*. It might instead indicate that the teachers did not give equal amounts of training to activities related to each subtest.

The finding that the IQ gains from Headstart were mostly on the non-*g* variance might explain why IQ gains from such programs fade with time (Brody, 1992). IQ tests given to people of different ages do not have the same items, as items that are useful for discriminating between small children are generally too easy for adults (Jensen, 1980). If one trains young children on the specific factors U_1, U_2, and U_3, and one later tests the same group with another test battery with the specific factors U_4, U_5, and U_6 then the earlier training would be irrelevant (barring any near-transfer effects), and therefore any IQ gain would vanish.

Alternatively, one might view the fading of IQ gains in light of the repeated finding that heritability increases with age, or equivalently, environmentality decreases with age (Plomin, DeFries, Knopik, & Neiderhiser, 2013). Since compensatory education is an environmental effect, its strength should decrease with time. As Jensen (1998, p. 184) pointed out, the most *g*-loaded subtests are also the most heritable ones, indicating that influencing *g* through environmental interventions is not easily accomplished.

4.1. Future studies

In this study, we focused on the Headstart program. Future studies should examine other compensatory educational programs to see whether the IQ gains were *g*-loaded. Indeed, the study of any phenomenon's relation to IQ scores could benefit from applying the method of correlated vectors. This is as true for compensatory education and dual n-back training (e.g., Jaeggi et al., 2010; but see Chooi & Thompson, 2012) as it is for fluoride poisoning (Choi, Sun, Zhang, & Grandjean, 2012) and myopia (Saw et al., 2004). The fact that the literature on intelligence focuses so much on manifest variables (i.e. IQ) leads to confusion in the press when phenomena such as the Flynn effect are reported, as well as when people observe that they can make their IQ scores go up by taking a test more than once. The only remedy is to focus on latent traits and always report the *g* loading.

A reviewer came up with an interesting suggestion for additional analyses, starting with the observation that the Headstart program is aimed at disadvantaged participants in the lower tail of the intelligence distribution, and therefore large and homogeneous *g* loadings are expected. So, it would be interesting to compare the *g* vector for these participants after the intervention with the vector of a comparable group without the intervention. The prediction for a successful program will be a significant reduction of the *g* loadings for the intervention group. The theoretical implication of such a result would be a reduction of the cognitive complexity of the completed measures.

4.2. Limitations of the studies

Estimates of the reliabilities of the vectors of *g* loadings were based on a very large number of high-quality studies. However, reliabilities of vectors of Headstart gains were based on a limited number of studies – albeit the complete empirical literature on this topic – leading to non-optimal estimates of the distribution of reliabilities.

4.3. Conclusion

Based on meta-analytical data and employment of the method of correlated vectors we showed that there is a strong and negative correlation of Headstart gains with *g*. Headstart programs can raise IQ test scores successfully, but not general mental ability per se. A very large amount of money was spent on Headstart programs, and it most likely led to increases in socially-desirable outcomes such as adequate nutrition, self-care skills, and social skills. However, our study shows it did not lead to the intended increase in intelligence as reflected in the very strong negative correlation with *g*. The outcomes of our study could be included in a cost–benefit analysis of Headstart programs.

Acknowledgment

We would like to thank Dr. John McNamara, Dr. Deborah Stipek, Dr. Emily Krohn, and Dr. Craig Ramey for their enthusiastic support of this project. We would also like to thank reviewers Roberto Colom and Meredith Frey and one anonymous reviewer for their constructive criticism of our paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.intell.2014.07.001.

References

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

Brody, N. (1992). *Intelligence*. New York: Academic Press.

Caruso, D. R., Taylor, J. J., & Detterman, D. K. (1982). Intelligence research and intelligent policy. In D. K. Detterman, & R. J. Sternberg (Eds.), *How and how much can intelligence be increased* (pp. 45–65). Norwood, NJ: Ablex Publishing Corporation.

Choi, A. L., Sun, G., Zhang, Y., & Grandjean, P. (2012). Developmental fluoride neurotoxicity: A systematic review and meta-analysis. *Environmental Health Perspectives*, *120*, 1362–1368.

Chooi, W. T., & Thompson, L. A. (2012). Working memory training does not improve intelligence in healthy young adults. *Intelligence*, *40*, 531–542.

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.

Eppig, C., Fincher, C. L., & Thornhill, R. (2010). Parasite prevalence and the worldwide distribution of cognitive ability. *Proceedings of the Royal Society B: Biological Sciences*, *277*(1701), 3801–3808.

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

Flynn, J. R., te Nijenhuis, J., & Metzen, D. (2014). The *g* beyond Spearman's *g*: Flynn's paradoxes resolved using four exploratory meta-analyses. *Intelligence*, *42*, 1–10.

*Gridley, B. E., Miller, G., Barke, C., Fischer, W., & Smith, D. (1990). Construct validity of the K-ABC with an at-risk preschool population. *Journal of School Psychology*, *28*, 39–49.

*Henderson, R. W., & Rankin, R. J. (1973). WPPSI reliability and predictive validity with disadvantaged Mexican-American children. *Journal of School Psychology*, *11*, 16–20.

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

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

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

Jaeggi, S. M., Studer-Luethi, B., Buschkuehl, M., Su, Y. F., Jonides, J., & Perrig, W. J. (2010). The relationship between n-back performance and matrix reasoning — Implications for training and transfer. *Intelligence*, *38*, 625–635.

Jensen, A. R. (1969). How much can we boost IQ and scholastic achievement. *Harvard Educational Review*, *39*, 1–123.

Jensen, A. R. (1980). *Bias in mental testing*. New York: Free Press.

Jensen, A. R. (1987). Differential psychology: Towards consensus. In S. Modgil, & C. Modgil (Eds.), *Arthur Jensen: Consensus and controversy*. New York: The Falmer Press.

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.

Johnson, W., Bouchard, T. J., Jr., Krueger, R. F., McGue, M., & Gottesman, I. I. (2004). Just one *g*: Consistent results from three test batteries. *Intelligence*, *32*, 95–107.

Johnson, W., te Nijenhuis, J., & Bouchard, T. J., Jr. (2008). Still just 1 *g*: Consistent results from five test batteries. *Intelligence*, *36*, 81–95.

*Krohn, E. J., Lamp, R. E., & Phelps, C. G. (1988). Validity of the K-ABC for a Black preschool population. *Psychology in the Schools*, *25*, 15–21.

*Lowe, J. D., Anderson, H. N., Williams, A., & Currie, B. B. (1987). Long-term predictive validity of the WPPSI and the WISC-R with Black school children. *Personality and Individual Differences*, *8*, 551–559.

Lynn, R. (2006). *Race differences in intelligence: An evolutionary analysis*. Augusta, GA: Washington Summit Publishers.

Lynn, R. (2013). Who discovered the Flynn effect? A review of early studies of the secular increase of intelligence. *Intelligence*, *41*, 765–769.

*McNamara, J. R., Porterfield, C. L., & Miller, L. E. (1969). The relationship of the Wechsler Preschool and Primary Scale of Intelligence with the coloured Progressive Matrices (1956) and the Bender Gestalt Test. *Journal of Clinical Psychology*, *25*, 65–68.

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

Nelson, G., Westhues, A., & MacLeod, J. (2003). A meta-analysis of longitudinal research on preschool prevention programs for children. *Prevention & Treatment*, *6*.

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

Plomin, R., DeFries, J. C., Knopik, V. S., & Neiderhiser, J. M. (2013). *Behavioral genetics* (6th ed.). New York: Worth.

Prokosch, M. D., Yeo, R. A., & Miller, G. F. (2005). Intelligence tests with higher *g*-loadings show higher correlations with body symmetry: Evidence for a general fitness factor mediated by developmental stability. *Intelligence*, *33*, 203–213.

Protzko, J., Aronson, J., & Blair, C. (2013). How to make a young child smarter: Evidence from the database of raising intelligence. *Perspectives on Psychological Science*, *8*, 25–40.

Ramey, C. T., Bryant, D. M., & Suarez, T. M. (1985). Preschool compensatory education and the modifiability of intelligence: A critical review. In D. Detterman (Ed.), *Current topics in human intelligence*. Norwood, NJ: Ablex Publishing Company.

Rushton, J. P. (1998). The "Jensen effect" and the "Spearman–Jensen hypothesis" of Black–White IQ differences. *Intelligence*, *26*, 217–225.

Rushton, J. P., & Ankney, C. D. (2009). Whole brain size and general mental ability: A review. *International Journal of Neuroscience*, *119*, 692–732.

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.

Saw, S. M., Tan, S. B., Fung, D., Chia, K. S., Koh, D., Tan, D. T., et al. (2004). IQ and the association with myopia in children. *Investigative Ophthalmology & Visual Science*, *45*, 2943–2948.

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

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

Somin, I. (2013). *Democracy and political ignorance: Why smaller government is smarter*. Stanford University Press.

Soriano, D., Duenas, M., & LeBlanc, P. (2006, August). The short-term and long-term effects of Head Start education and no child left behind. *Paper presented at the meeting of the Association of Teacher Educators, Philadelphia, PA*.

Spitz, H. H. (1986). *The raising of intelligence: A selected history of attempts to raise retarded intelligence*. Hillsdale, NJ: Erlbaum.

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., & van der Flier, H. (2007). Score gains on *g*-loaded tests: No *g*. *Intelligence*, *35*, 283–300.

Thompson, J. (2013). The Flynn effect re-evaluated. In James Thompson (Ed.), *Intelligence. Vol. 41, Issue 6.* (pp. 751–858) (November–December 2013).

U.S. Department of Health and Human Services (2012). Head Start program facts fiscal year 2012. http://eclkc.ohs.acf.hhs.gov/hslc/mr/factsheets/2012-hs-program-factsheet.html (Retrieved from)

van Haasen, P. P., de Bruyn, E. E. J., Pijl, Y. J., Poortinga, Y. H., Lutje Spelberg, H. C. , Vander Steene, G., et al. (1986). *WISC-R: Wechsler Intelligence Scale for Children — Revised; Nederlandstalige uitgave [WISC-R: Wechsler Intelligence Scale for Children — Revised; Dutch edition]*. Lisse, The Netherlands: Swets.

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., Must, O., & Must, A. (2014). Controlling for increased guessing enhances the independence of the Flynn effect from *g*: The return of the Brand effect. *Intelligence*, *43*, 27–34.

*Yater, A. C., Barclay, A., & Leskosky, R. (1971). Goodenough–Harris drawing test and WPPSI performance of disadvantaged preschool children. *Perceptual and Motor Skills*, *33*, 967–970.

Zill, N., Resnick, G., Kim, K., O'Donnell, K., & Sorongon, A. (2003). *Head Start faces 2003: A whole-child perspective on program performance*. Administration for Children and Families, U.S. Department of Health and Human Services (Retrieved December 1, 2005, from: http://www.acf.hhs.gov/programs/opre/hs/faces/reports/faces00_4thprogress/faces 00_4thprogress.pdf).