Is the Flynn Effect Primarily a Rise in Structural Visualization?

Christopher A. Condon

and

David H. Schroeder

JOHNSON O'CONNOR RESEARCH FOUNDATION, INC.

Technical Report 2008-1

April 2008

Is the Flynn Effect Primarily a Rise in Structural Visualization?

Christopher A. Condon and David H. Schroeder

ABSTRACT

An extensive literature has documented the steady rise across time in performance on IQ tests, generally known as the Flynn effect (Neisser, 1998). While attention has been focused on IQ and other indexes of general performance, there has been only limited investigation of patterns for more-specific variables. In this report we use data on specific aptitude measures to demonstrate a rise across time in performance on visuospatial tests and to evaluate whether the Flynn effect in these data is general across tests or largely specific to spatial tests.

In particular, we examined scores for the Foundation's tests for three points in time, 1989-90, 1995-96, and 2002-03. To facilitate comparison with the IQ studies, we formed an IQ-like index, or "general aptitude score," by summing scores on nine Foundation tests. We also summed scores for four more-specific domains, based on factors identified previously in the Foundation's battery: Reasoning, Numerical, Spatial, and Memory. We found an increase of .07 standard-deviation (*SD*) units from 1989-90 to 2002-03 in scores on the general aptitude index, which confirms that there is an upward overall trend on our tests. For the domains, we found a relatively large increase of .18 *SD* units in the spatial area, a smaller gain of .08 *SD* units for memory, a negligible gain for numerical performance (.01), and a small loss for reasoning (-.09).

We addressed whether or not the changes in scores across time have altered the magnitudes of sex differences on our tests and found that sex differences have generally remained the same. Similarly, we examined whether the patterns of age differences on Foundation tests have changed and found that for the most part, they have stayed the same.

Finally, we investigated the specific hypothesis that the gain in the general aptitude index can be attributed to the more-specific gain in the spatial domain. We formed a nonspatial general aptitude index by summing the seven nonspatial test scores and found that scores on this index increased by only .01 *SD* units over the 13-year interval.

Thus, it appears that the rise in performance that we observed on our tests is mostly in the spatial domain—that is, in structural visualization. It could be the case that changes in the culture, such as the greater use of visual and graphical representations of information (on computers and elsewhere) are leading to greater development of structural visualization, while general increases in aptitude may be much smaller than some observers have thought.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGMENTS

This technical report is based on a paper presented by the authors at the eighth annual convention of the International Society for Intelligence Research in December 2007 (Condon & Schroeder, 2007). We have also drawn from an earlier paper presented to the same group (Condon & Schroeder, 2002).

The conduct of this study was made possible by the creation and maintenance of the Research Department's bargraph database over the last several decades. This database has been the product of the Foundation's testing staff, who collected the data, and the Research Department's professional and support staff over the years, and the authors wish to acknowledge their contributions.

The work reported here has also received the support of the leadership of the Foundation. In particular, we acknowledge Mr. David Ransom, President of the Foundation, and Mr. Robert Kyle, Vice President of the Foundation, for seeing value in the identification of changes in aptitude performance over time.

INTRODUCTION

An investigator named James Flynn at the University of Otago, New Zealand, has attracted quite a bit of attention with his finding that performance on IQ tests in industrialized countries has risen substantially over the 20th century (Neisser, 1998). He estimated that performance in the United States increased by just over one standard deviation (*SD*) between 1947-48 and 2001-02, which corresponds to over 15 IQ points (Flynn, 2007, pp. 8, 180-181). This result is shown in Figure 1 (the lower of the two lines), in which 1947-48 performance is set at zero and the later years are shown relative to 1947-48, in IQ points. This specific finding is based on a comparison of the performance of U.S. standardization samples for the Wechsler Intelligence Scale for Children (WISC) in 1947-48, 1972, 1989, and 2001-02. Flynn has found similar patterns for the Wechsler Adult Intelligence Scale and the Stanford-Binet intelligence test and for other nations such as Great Britain, The Netherlands, and Norway (Flynn, 2007).

Flynn has also studied performance on Raven's Progressive Matrices, which is a measure of abstract reasoning and correlates highly with scores on IQ tests. He estimated a gain of almost two *SD*s for the Raven's test over the period from 1947-48 to 2001-02 (see Figure 1, the upper of the two lines).

The Flynn effect has been interpreted by some observers as an increase in *g*, which is thought to be a unitary general factor related to performance across cognitive-ability tests (Anastasi & Urbina, 1997, pp. 310-311). Flynn has recently interpreted the effect, however, in terms of changes in specific emphases in society (Flynn, 2007). Flynn noted that the gains on individual subtests within the Wechsler tests are quite variable. For example, on the WISC during the period from 1947-48 to 2001-02, performance on the Similarities subtest increased by the equivalent of 24 IQ points (1.6 standard-deviation units), while performance on the Information subtest went up only two points.[1] Flynn argued that the reason that scores on subtests such as Similarities have risen so much is because in the modern world, people are trained to think in abstract, scientific ways rather than more-concrete, prescientific ways, and this kind of thinking is rewarded on tests like Similarities.

[1] In the Similarities subtest, examinees are asked to identify in what way two things are alike (e.g., a car and a train). On Information, examinees are asked questions about general information common across a culture, such as "How many weeks are there in a year?".

It can also be noted that the world today presents persons with more visual information than in years past, and so one could also consider whether this is having an effect on structural visualization. In this report we will examine trends in specific abilities, including structural visualization, and consider the effect of these trends on estimates of composite, or "general," aptitude.[2]

The Flynn effect is of particular interest to persons working in aptitude measurement for two reasons: (a) it poses important questions regarding why performance improves across cohorts—that is, questions about genes, environment, and the sources of individual differences, and (b) on a practical level, it speaks to the need to construct new norms periodically in order to stay in line with the rising levels of performance. For example, if performance on a given test were rising at the same rate as IQ scores (.2 standard-deviation units per decade), then over a 20-year span, scores would increase by .4 *SD* units, and a score at the 50th percentile at the beginning of the period would represent only the 34th percentile by the end of the period.

METHOD

Samples

The examinees for this study were unselected paying clients of the Foundation tested in the 11 current Foundation offices, as well as Tampa and New Orleans. Examinees took the Foundation battery generally for the purpose of gaining information about their aptitudes that they might use in making academic and occupational decisions. For this study, we used examinees from 1989-90, 1995-96, and 2002-03 to examine score trends over time. By using these time periods, we were able to consider the widest possible range of times while also using test scores that could be placed on common scales.[3]

The first sample consisted of 10,969 examinees tested in 1989 and 1990, including 5,754 (52.5%) males and 5,215 (47.5%) females. Examinees ranged in age from 14 to 70 (M = 27.38, SD = 10.15). The second sample consisted of 8,092

[2] Outside research indicates that composite scores from tests such as those of the Foundation can reasonably be compared to composite scores on instruments such as the Wechsler tests (Carroll, 1993).

[3] For example, we did not use data from 1992 because the form of Analytical Reasoning used at that time, Wks. 244 F, could not be equated with the form used later (244 I).

examinees tested in 1995 and 1996, with 4,215 (52.1%) males and 3,877 (47.9%) females. Examinees ranged in age from 14 to 73 (M = 27.26, SD = 10.60). Finally, the third sample consisted of 6,703 examinees tested in 2002 and 2003, including 3,595 (53.6%) males and 3,108 (46.4%) females. Examinees ranged in age from 14 to 72 (M = 25.33, SD = 10.56). The characteristics of the Foundation's testing population are described more fully in Statistical Bulletin 1998-3.

Measures

The examinees used in this study took the Foundation's full standard battery, and for this study we analyzed scores on nine cognitive tests in the battery: Analytical Reasoning, Inductive Reasoning, Number Facility, Number Series, Paper Folding, Memory for Design, Observation, Number Memory, and Silograms.[4] For further information on each of the nine tests, including the reliability and the aptitude that the test measures, see Table 1.

For each test, we standardized the raw scores so that the test scores were on the same scale. We also partialled age from the scores because we wanted to eliminate the possibility that different age distributions for the three samples might create spurious differences between the groups. (The exception here was for the section on age differences, for which we used unpartialled scores.)

For this study, we examined scores at three levels of generality: individual tests, domains (such as memory), and overall performance. The overall score, which we termed an index of "general aptitude," was formed by summing the standardized scores for the nine tests. The resulting score allowed us to compare our results with previous studies that used overall indexes (usually IQ) that combined scores across a number of different tasks. In creating this index, we viewed it as a statistical composite that does not necessarily correspond to a single entity in the natural world.

Among the three levels, we emphasized the domain level in this study. This level allowed us to show the differing patterns of change across time that were obscured at the general aptitude level. We did not focus on individual tests here, although we report the results for them in appendixes. Those findings

[4] We did not use scores for four of the Foundation's cognitive tests. For Wiggly Block, there were form changes in the 1990s such that we did not wish to combine raw scores for comparison. For Number Checking, Ideaphoria, and English Vocabulary, the tests could not be grouped into common domains, such as Memory, which we needed to do for this conference presentation (see the remainder of the Method section; Condon & Schroeder, 2007).

could form the basis for a subsequent report, although it might be helpful to include more years of data to ensure reliable results.

We formed domain scores for four areas, corresponding to the four factors that we previously identified in the Foundation's battery (Condon & Schroeder, 2003): Spatial, Memory, Numerical, and Reasoning. For the Spatial domain, we summed standardized scores for Paper Folding and Memory for Design. As noted in Footnote 4, we could not use scores on Wiggly Block because of form changes in the late 1990s. In order to have at least two test scores for the domain, we used scores for Memory for Design because our 2003 factor analysis had indicated that, although Memory for Design has historically been thought of as a memory test, it loads substantially on both the Spatial and Memory factors.

For the Memory domain, we summed standardized scores on Silograms, Number Memory, and Observation. For Numerical, we used scores on Number Series and Number Facility. Finally, for Reasoning, we used scores on Inductive Reasoning and Analytical Reasoning.

For both the domain and general aptitude scores, after summing standardized scores for the relevant individual tests, we restandardized so that scores on the resulting variables were on common scales, with standard deviations equal to one.

RESULTS

Overall Sample

General aptitude score. In Table 2 and Figure 2 we show the changes in the mean general aptitude score from 1989-90 to 1995-96 and 2002-03. As can be seen, we set the 1989-90 mean at zero so that the later values can be viewed as mean changes relative to 1989-90. There is a modest upward trend of .07 *SD* units for the general aptitude score from 1989-90 to 2002-03. Thus, there is a modest Flynn effect for the Foundation battery, but the magnitude of the effect is less than what Flynn observed for IQ tests.

Four domains. In addition to looking at the general aptitude score, we examined the four domain scores, as shown in Table 2 and Figure 3. As can be seen, the four domain areas, which together make up the general aptitude score,

show very different patterns of change. The spatial domain showed the largest gain, .18 *SD* units from 1989-90 to 2002-03. Performance in the memory domain also increased but at roughly half of the rate of the spatial domain, .08 *SD* units. The numerical domain showed essentially no change across the years, while the reasoning domain showed a modest decrease. To compare the four domains, we performed a one-way within-subjects ANOVA on the 2002-03 means. The ANOVA showed significant differences among the four domains in the mean change values, $F(3, 6702) = 149.78$, $p < .001$. Having demonstrated significant differences among the domains with the omnibus ANOVA, we examined the domain means in a pairwise manner. First, we performed a within-subjects *t*-test on the two closest means, namely, the means for the numerical and memory domains. Since that difference was significant, $t(6702) = 5.79$, $p < .001$, we can infer that every other pairwise comparison among the four means is also significant.

Individual tests. In addition to examining the domain differences, we also computed the changes across time for the nine individual tests, and these values are shown in Appendix A.

To reiterate, the analyses for the overall sample indicate that the four domains show widely varying patterns of changes over time, with the largest rate of gain for the spatial domain.

Sex Differences

Some investigators have reported that sex differences in cognitive abilities are declining across time (Hyde, Fennema, & Lamon, 1990), although this finding has not replicated consistently (Hedges & Nowell, 1995). If sex differences are becoming smaller, this implies that the trends across time are different for males and females. For example, on structural visualization tests, this would imply that females are gaining faster than males. In our next analysis, we investigated this possibility in our data.

General aptitude score. In Table 3 and Figure 4, we show the trends for males and females on the general aptitude score from 1989-90 to 2002-03. As in the earlier analysis, we set the 1989-90 means to zero so that the other values show mean changes relative to 1989-90. As can be seen, the trend lines for males and females are quite similar to each other and to the trend for the overall sample (see earlier figure). As of 2002-03, the mean for females had increased by .07 *SD* units, while the mean for males had increased .06 *SD* units.

Four domains. Next, we looked at the trends across the four domains for females and males. First, in Table 3 and Figure 5 we show the changes in performance for females for the four domains from 1989-90 to 2002-03. In general, the pattern is very similar to the pattern for the overall sample. Females showed the largest gain in performance for the spatial domain (.19 *SD* units), followed by memory (.10), and, negligibly, numerical (.01). For reasoning, they showed a decrease across the time span (-.09). A within-subjects one-way ANOVA showed significant differences among the 2002-03 change means for the four domains, $F(3, 3107) = 84.22$, $p < .001$. A within-subjects *t*-test comparing the two closest means, numerical and reasoning, showed that the difference between them is significant, $t(3107) = 5.01$, $p < .001$. Therefore, all other differences between the 2002-03 change means for domains are significant as well.

The trends for males are shown in Table 3 and Figure 6. The pattern for males is very similar to that for females, with the largest increase being in the spatial domain (.16 *SD* units), followed by memory (.08) and numerical (.01), with reasoning performance showing a decrease (-.09). A within-subjects one-way ANOVA showed significant differences among the four domains, $F(3, 3594) = 68.27$, $p < .001$. A within-subjects *t*-test comparing the two closest means, numerical and memory, showed a significant difference, $t(3594) = 3.85$, $p < .001$. Therefore, all other differences between domains are significant as well.

Thus, the patterns of changes across time for males and females on general aptitude and the four domains are similar to each other and to the patterns for the overall sample. Consequently, one would not expect to see much change in the size of the differences between males and females, and this expectation is borne out in our data set. In Table 4 and Figure 7 we show the sex differences for the four domains for the three points in time.[5] As can be seen, the magnitudes of the sex differences show little change across the three time periods.

Individual tests. We also computed the changes across time by sex for the nine individual tests, and those means are shown in Appendix B.

[5] It may be noted that in Table 3 and Figures 4-6, we set the 1989-90 means for males and females to zero in order to show the amount of change from 1989-90 to 1995-96 and 2002-03. In Table 4 and Figure 7, we compare male and female performance to the same frame of reference so that the mean differences between the sexes can be seen.

Age Differences

As discussed earlier, Flynn's research has shown a fairly steady increase in performance over a number of decades. Nonetheless, there have been some indications of changes in the rate of gain, especially in recent years (Teasdale & Owen, 2008). To the extent that different cohorts show somewhat different rates of change, patterns of age differences can be expected to shift as the cohorts showing greater and lesser gains move through the lifespan. To evaluate this possibility, we compared different age groups for the three time periods in the study: 1989-90, 1995-96, and 2002-03.

General aptitude score. In Table 5 and Figure 8 we show the trends for four age groups (14-20, 21-28, 29-36, 37+) on the general aptitude score from 1989-90 to 2002-03. As noted earlier, the data here were not partialled for age. However, the means for 1989-90 for each age group were set to zero so that the values for 1994-95 and 2002-03 represent changes relative to 1989-90. As can be seen, the trend lines for the age groups are quite similar to each other. In fact, a one-way ANOVA showed that the four age groups did not significantly differ in their 2002-03 change means, $F(3, 6699) = .344$, $p = .794$.

Four domains. Next, we looked at the trends across the four domains for the four age groups. In Figure 9 we show the changes in performance for the four age groups for the spatial domain from 1989-90 to 2002-03, while in Table 5 we display the values for all four domains. As shown in the figure, there were sizable increases in the spatial domain for the 14-20, 21-28, and 29-36 age groups, and a more-modest increase for the 37+ age group. A one-way ANOVA showed significant differences in the 2002-03 change means for the spatial domain, $F(3, 6699) = 5.81$, $p = .001$. The only other domain that showed significant differences among the 2002-03 change means was the numerical domain, $F(3, 6699) = 6.63$, $p < .001$.

Individual tests. We also computed the changes across time by age for the nine individual tests, and those means are shown in Appendix C.

In general, the results for the four age groups mirror those of the overall sample shown at the beginning of the Results section, with the largest gains observed for the spatial domain. As a consequence, as with the sex-difference analysis, the differences among the age groups remained about the same over the time period of the study.

Latent-Variable Analysis

In our analyses to this point, we used simple sums of standardized raw scores for our tests. This facilitated comparison with other Flynn-effect studies, most of which used IQ tests, in which the overall score is essentially a sum of the subtest scores. We were also interested, however, in analyzing our data in terms of latent variables (that is, factors) because such an analysis could be expected to shed light on the issue of which variables are actually showing gains over time (specifically, general aptitude versus the spatial domain). We describe those analyses in Appendix D. Unfortunately, the analyses could not be completed because in the model that we used, the individual tests were expected to show gains in proportion to their loadings on the respective factors. Inasmuch as our tests showed somewhat different patterns of gains within factors (e.g., within the Memory factor, Observation showed larger gains than Silograms and Number Memory), this model did not fit our data in an acceptable way, and we could not pursue the analyses further. There may be an alternative model in which specific test variance for individual tests is allowed to account for differing patterns of gain within factors.

Spatial Versus General Aptitude Gains

Finally, using raw-score-based methods, we return to the issue of whether the gains in the general aptitude index can be accounted for by the gains in the spatial domain. As previously discussed and shown in Table 6 and Figure 10, the gains over the 13-year interval of the study were .07 *SD* units for the general aptitude index and .18 units for the spatial domain. We formed a nonspatial general aptitude index by summing scores on the seven tests in the data set excluding the two spatial tests. As can be seen, for this index the difference between the 1989-90 and 2002-03 means was only .01 *SD* units. This finding indicates to us that outside of the spatial area, our data do not show a general rise in performance.[6] It may be the case that the widespread claim that the Flynn effect is a general phenomenon is misdirected, and the effect is actually a spatial phenomenon. The Foundation's historical attention to specific aptitudes has allowed us to identify a pattern that was obscured by the focus on IQ tests.

[6] This is not to overlook the modest gain in performance in the area of memory, which counteracted the small decline for reasoning.

DISCUSSION

Thus, as we have noted, the rise in performance over time on Foundation tests is primarily in the area of spatial-related tests. Although the Flynn effect is widely viewed as an increase in a general dimension of performance, the gains in our samples were only in specific areas. It will be interesting to see if other investigators find similar results (see Cocodia et al., 2003).

One existing body of outside data that permits examination of specific variables is the data reported by Flynn for the Wechsler Intelligence Scale for Children (WISC; Flynn, 2007). Flynn found sizable gains over time for two spatial-related subtests, Block Design and Object Assembly. The WISC does not have a conventional mental-rotation test such as the Foundation's Paper Folding test or Thurstone's Flags test (Thurstone & Thurstone, 1941, p. 61), and we conjecture that such tests might have shown more-striking gains than did the two WISC tests. In fairness, it should also be noted that the WISC data shows the largest gains for the Similarities test, and that effect appears to be distinct from the effect for the spatial area.

In addition, we noted earlier that the gains on Raven's Matrices have been consistently larger than the gains in IQ (Flynn, 2007), and this could be because the Raven's test uses figural images (and hence visual processing) rather than words or numbers.

An intriguing issue related to patterns across time is why performance (in our data, in the spatial domain) seems to be increasing. Other investigators have offered a number of suggestions (Neisser, 1998). Some have pointed to nonsubstantive variables such as testwiseness (see Flynn, 1998, p. 42), which in all likelihood has increased relative to, say, 70 years ago. Nonetheless, the majority of investigators believe there has been at least some substantive gain beyond the gain due to influences such as testwiseness in more-recent decades.

In terms of the Foundation's tests, in-house research has indicated that structural visualization has a heritability estimate of .60 (Schroeder, Johnson, & Nakajima, 1996). This means that structural visualization has a substantial genetic component, but it has some environmental influence also. Other Foundation research has shown that individual differences in structural visualization are highly stable in adulthood (Technical Report 1997-1). This

implies that the environmental influences on structural visualization take place early in one's life and do not change one's standing at later ages.

Inasmuch as contemporary examinees score higher on structural visualization than examinees in the 1980s, as our research indicates, this gain is likely to be due to the environmental component of visualization. (It is unlikely that the genetic stock for visualization has changed substantially over this time.) How might this have happened? Other observers have noted that we live in an increasingly visual world, with graphical interfaces on computers, PowerPoint slides in lectures, and Wii consoles at home. It may be the case that genetically-based potential has remained steady, but environmental influences during childhood are now more favorable for the development of structural visualization, and so persons are realizing a greater proportion of their potential. Alternatively, it could be the case that the contemporary culture is boosting performance via a "practice" effect, but the research literature on practice indicates that it is only efficacious when the practice is intense or the task is highly similar to the task on the test (Brody, 1992).

Some other points should be made:

1. This study illustrates the point that has been emphasized within the Foundation that many phenomena in the area of cognitive tests that others think of as "general" are actually specific to particular areas or aptitudes, and examining performance at a more-specific level allows us to see that.
2. The decline in performance observed for the reasoning domain, although small and somewhat erratic (Statistical Bulletin 1998-3), is curious and merits further investigation.
3. As noted earlier, when performance is changing over time, it can be important to update norms so that percentiles are appropriate for current examinees.
4. It should be borne in mind that these findings are derived from the Foundation's tests administered to the Foundation's testing population. Hence other research will be needed to show whether other measures and other populations yield the same pattern of results. Also, since our study addressed the period from 1989-90 to 2002-03, it is possible that performance gains in other time periods will show different relationships.

REFERENCES

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Brody, N. (1992). *Intelligence* (2nd ed.). London: Academic Press.

Byrne, B. M. (2001). *Structural equation modeling with AMOS.* Mahwah, NJ: Lawrence Erlbaum Associates.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge, UK: Cambridge University Press.

Cocodia, E. A., Kim, J.-S., Shin, H.-S., Kim, J.-W., Ee, J., Wee, M. S. W., et al. (2003). Evidence that rising population intelligence is impacting in formal education. *Personality and Individual Differences*, *35*, 797-810.

Condon, C. A., & Schroeder, D. H. (2002, December). *Is the Flynn effect slowing down? An examination of recent data for* g *and specific abilities*. Paper presented at the annual meeting of the International Society for Intelligence Research, Nashville, TN.

Condon, C. A., & Schroeder, D. H. (2003, August). *An SEM analysis of the Johnson O'Connor Research Foundation's battery*. Poster session presented at the annual meeting of the American Psychological Association, Toronto.

Condon, C. A., & Schroeder, D. H. (2007, December). *Is the Flynn effect primarily a rise in spatial ability?* Paper presented at the annual meeting of the International Society for Intelligence Research, Amsterdam.

Dolan, C.V., Colom, R., Abad, F.J., Wicherts, J. M., Hessen, D. J., & van de Sluis, S. (2006). Multi-group covariance and mean structure modeling of the relationship between the WAIS-III common factors and sex and educational attainment in Spain. *Intelligence, 34(2),* 193-210.

Flynn, J. R. (1998). IQ gains over time: Toward finding the causes. In U. Neisser (Ed.), *The rising curve: Long-term gains in IQ and related measures* (pp. 25-66). Washington, DC: American Psychological Association.

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

Hedges, L. V., & Nowell, A. (1995). Sex differences in mental scores, variability, and numbers of high-scoring individuals. *Science*, *269*, 41-45.

Hyde, J. S., Fennema, E., & Lamon, S. J. (1990). Gender differences in mathematics performance: A meta-analysis. *Psychological Bulletin*, *107*, 139-155.

Neisser, U. (1998). *The rising curve: Long-term gains in IQ and related measures*. Washington, DC: American Psychological Association.

Schroeder, D. H., Johnson, W. G., & Nakajima, M. (1996, July). *Understanding genetic factors in cognitive abilities: One general factor, specific factors, or group factors?* Paper presented at the annual meeting of the American Psychological Society, San Francisco.

Statistical Bulletin 1998-3. *Statistical summary of standard test battery population, 1984-1996*. L. L. Meyer & D. H. Schroeder. Chicago: Johnson O'Connor Research Foundation.

Teasdale, T. W., & Owen, D. R. (2008). Secular declines in cognitive test scores: A reversal of the Flynn Effect. *Intelligence*, *36*, 121-126.

Technical Report 1997-1. *Long-term stability of 11 aptitude tests*. J. K. Bethscheider & D. H. Schroeder. Chicago: Johnson O'Connor Research Foundation.

Thurstone, L. L., & Thurstone, T. G. (1941). Factorial studies of intelligence. *Psychometric Monographs* (No. 2). Chicago: University of Chicago.

Table 1

Nine Foundation Tests Used in the Current Study

Test	Reliability	Aptitude measured
Inductive Reasoning	.84	Quickness in seeing relationships among separate facts, ideas, or observations.
Analytical Reasoning	.81	Ability to arrange ideas into a logical sequence.
Number Series	.87	Ability to reason (solve problems) with numbers.
Number Facility	.86	Ability to perform arithmetic operations quickly and accurately.
Paper Folding	.82	Visualizing in 3-D.
Memory for Design	.80	Memory for straight-line patterns.
Silograms	.92	Associative memory for verbal material.
Number Memory	.82	Memory for numbers.
Observation	.62	Ability to retain a mental image of various objects in the mind and quickly perceive any changes in the nature or position of an object.

Table 2

Changes in Mean Standardized Scores From 1989-90 to 2002-03

Score	Year-group		
	1989-90	1995-96	2002-03
Overall			
General aptitude	.00	.06	.07
Domain			
Spatial	.00	.12	.18
Memory	.00	.08	.08
Numerical	.00	.01	.01
Reasoning	.00	-.06	-.09

Note. *N* for 1989-90 = 10,969, *N* for 1995-96 = 8,092, *N* for 2002-03 = 6,703. General aptitude = sum of nine tests in the study; Spatial = sum of Paper Folding and Memory for Design; Memory = sum of Silograms, Number Memory, and Observation; Numerical = sum of Number Series and Number Facility; Reasoning = sum of Inductive Reasoning and Analytical Reasoning. The values for 1989-90 were set to zero so that the other values represent mean changes relative to 1989-90. For 2002-03, the four domain means all differed significantly from each other.

Table 3

Changes in Mean Standardized Scores From 1989-90 to 2002-03 by Sex

	Sex					
	Male			Female		
Score	1989-90	1995-96	2002-03	1989-90	1995-96	2002-03
Overall						
General aptitude	.00	.03	.06	.00	.08	.07
Domain						
Spatial	.00	.11	.16	.00	.14	.19
Memory	.00	.06	.08	.00	.11	.10
Numerical	.00	.00	.01	.00	.02	.01
Reasoning	.00	-.08	-.09	.00	-.04	-.09

Note. Male: *N* for 1989-90 = 5,754, *N* for 1995-96 = 4,215, *N* for 2002-03 = 3,595. Female: *N* for 1989-90 = 5,215, *N* for 1995-96 = 3,877, *N* for 2002-03 = 3,108. General aptitude = sum of nine tests in the study; Spatial = sum of Paper Folding and Memory for Design; Memory = sum of Silograms, Number Memory, and Observation; Numerical = sum of Number Series and Number Facility; Reasoning = sum of Inductive Reasoning and Analytical Reasoning. The values for 1989-90 were set to zero so that the other values represent mean changes relative to 1989-90. For 2002-03, within each sex, all four domain means differed significantly from each other.

Table 4

Differences Between Males and Females by Domains From 1989-90 to 2002-03

	Difference score (*d*)		
Domain	1989-90	1995-96	2002-03
Spatial	.30	.26	.25
Memory	-.36.	-.41	-.38
Numerical	.02	.00	.03
Reasoning	-.11	-.16	-.10

Note. *N* for 1989-90 = 10,969, *N* for 1995-96 = 8,092, *N* for 2002-03 = 6,703. Spatial = sum of Paper Folding and Memory for Design; Memory = sum of Silograms, Number Memory, and Observation; Numerical = sum of Number Series and Number Facility; Reasoning = sum of Inductive Reasoning and Analytical Reasoning. The difference score is the statistic *d*, which indicates the difference between male and female performance in standard-deviation units. Positive values are in the direction of higher male performance.

Table 5

Changes in Mean Standardized Scores From 1989-90 to 2002-03 by Age Group

Score	Year-group 1989-90	1995-96	2002-03
Overall			
General aptitude			
14-20 yrs.	.00	.06	.04
21-28	.00	-.01	.04
29-36	.00	.06	.06
37+	.00	.09	.08
Domain			
Spatial			
14-20	.00	.16	.21
21-28	.00	.08	.18
29-36	.00	.12	.18
37+	.00	.08	.07
Memory			
14-20	.00	.07	.04
21-28	.00	.05	.11
29-36	.00	.08	.06
37+	.00	.10	.09
Numerical			
14-20	.00	.02	-.02
21-28	.00	-.08	-.05
29-36	.00	.04	.00
37+	.00	.08	.11
Reasoning			
14-20	.00	-.09	-.12
21-28	.00	-.09	-.13
29-36	.00	-.05	-.07
37+	.00	-.01	-.06

Note. *N*s for 14-20: 1989-90 = 3,787; 1995-96 = 2,947; 2002-03 = 3,174. *N*s for 21-28: 1989-90 = 3,061; 1995-96 = 2,263; 2002-03 = 1,665. *N*s for 29-36: 1989-90 = 1,840; 1995-96 = 1,211; 2002-03 = 759. *N*s for 37+: 1989-90 = 2,281; 1995-96 = 1,671; 2002-03 = 1,105. General aptitude = sum of nine tests in the study; Spatial = sum of Paper Folding and Memory for Design; Memory = sum of Silograms, Number Memory, and Observation; Numerical = sum of Number Series and Number Facility; Reasoning = sum of Inductive Reasoning and Analytical Reasoning. The values for 1989-90 were set to zero so that the other values represent mean changes relative to 1989-90. The general aptitude means for 2002-03 did not significantly differ from each other. Regarding the domains, there were significant differences among the means for 2002-03 for Spatial and Numerical, respectively, and not among the means for Memory and Reasoning.

Table 6

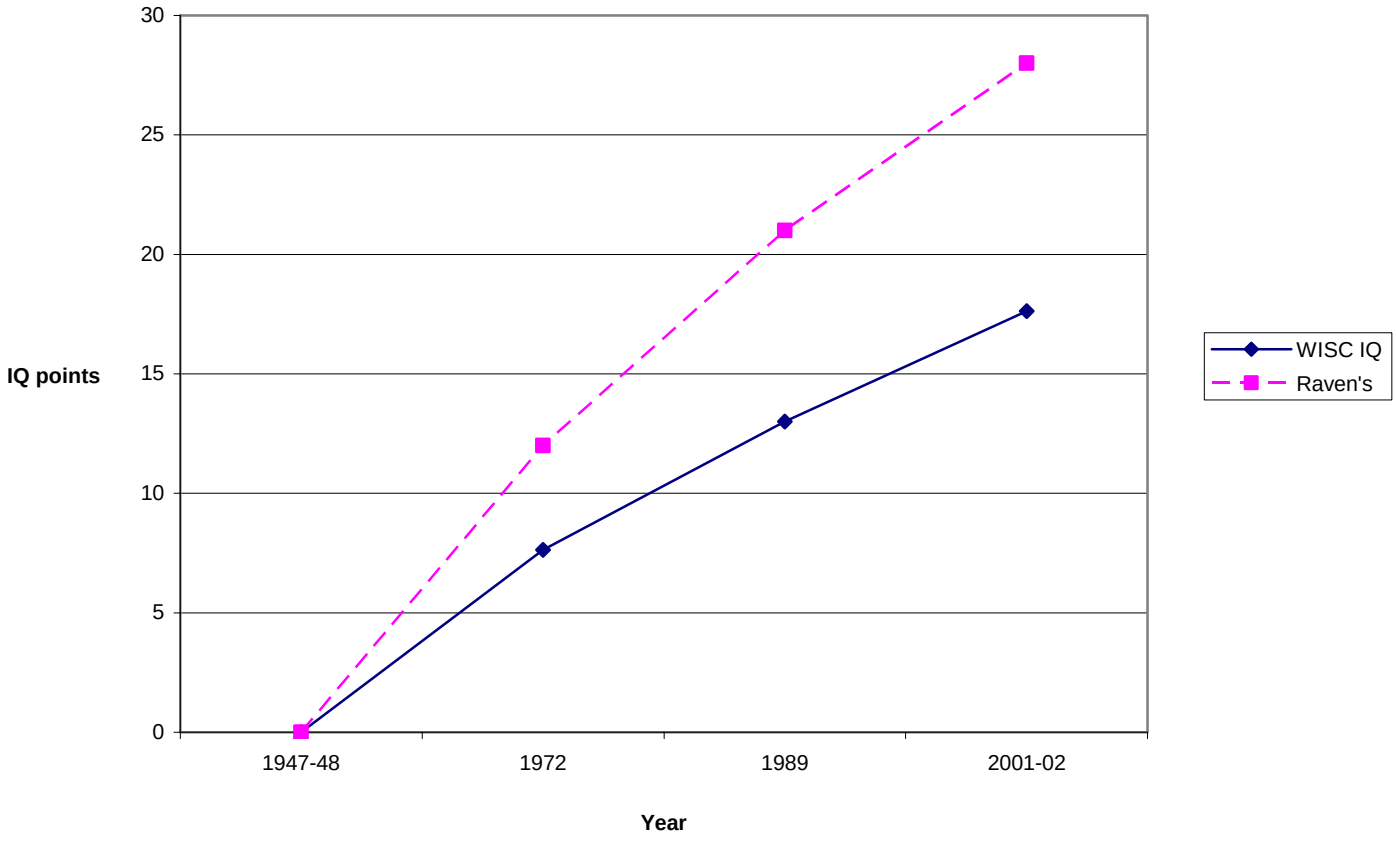
Changes in Mean Standardized Scores for General Versus Spatial and Nonspatial General Aptitude From 1989-90 to 2002-03

	Year-group		
Score	1989-90	1995-96	2002-03
General aptitude	.00	.06	.07
Spatial domain	.00	.12	.18
Nonspatial general aptitude	.00	.02	.01

Note. *N* for 1989-90 = 10,969, *N* for 1995-96 = 8,092, *N* for 2002-03 = 6,703. General aptitude = sum of nine tests in the study; spatial domain = sum of Paper Folding and Memory for Design; nonspatial general aptitude = sum of seven tests not including Paper Folding and Memory for Design. The values for 1989-90 were set to zero so that the other values represent mean changes relative to 1989-90. For 2002-03, the three means all differed significantly from each other.

Figure 1

Gains in Performance on the WISC and Raven's Matrices From 1947-48 to 2001-02



Note. The values in this figure are based on representative U.S. samples.

Figure 2

Changes in Mean General Aptitude Score From 1989-90 to 2002-03

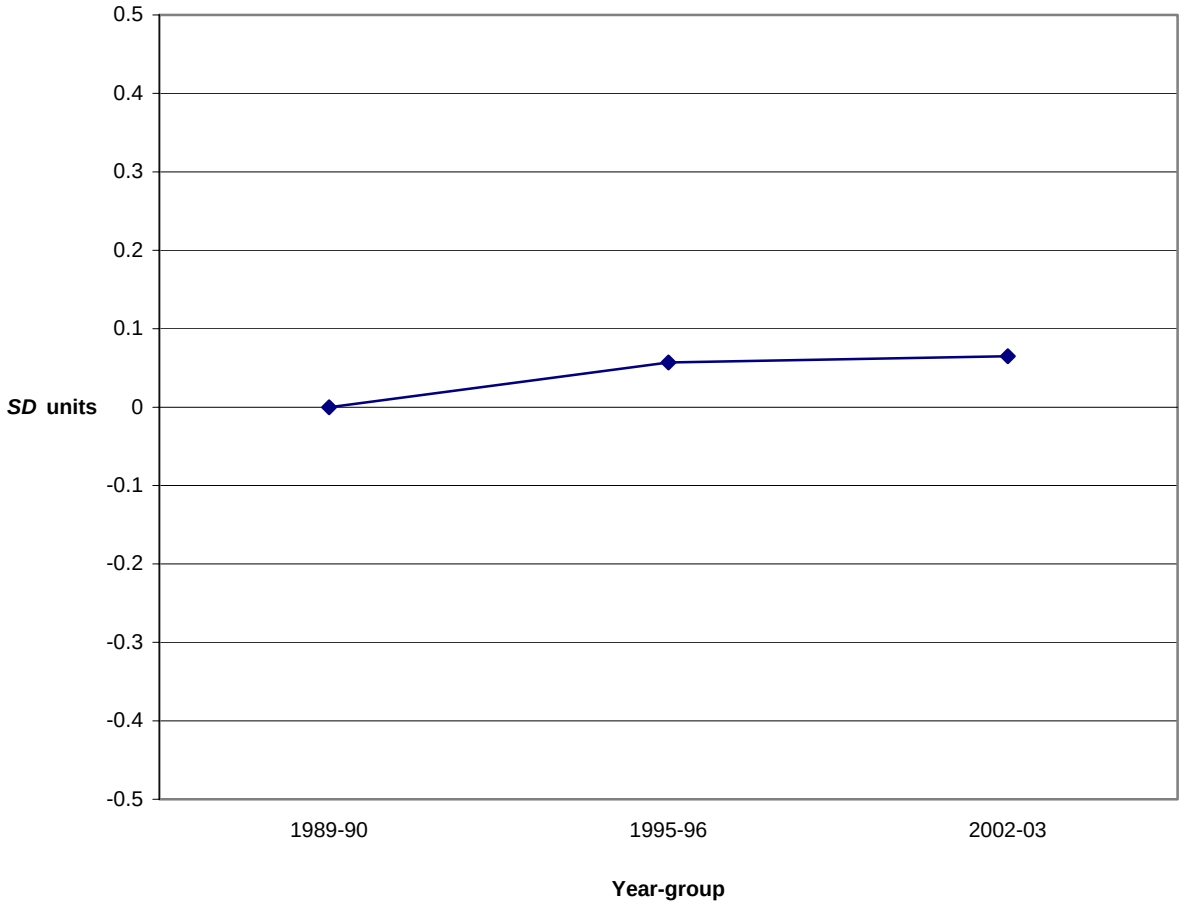


Figure 3

Changes in Mean Domain Scores From 1989-90 to 2002-03

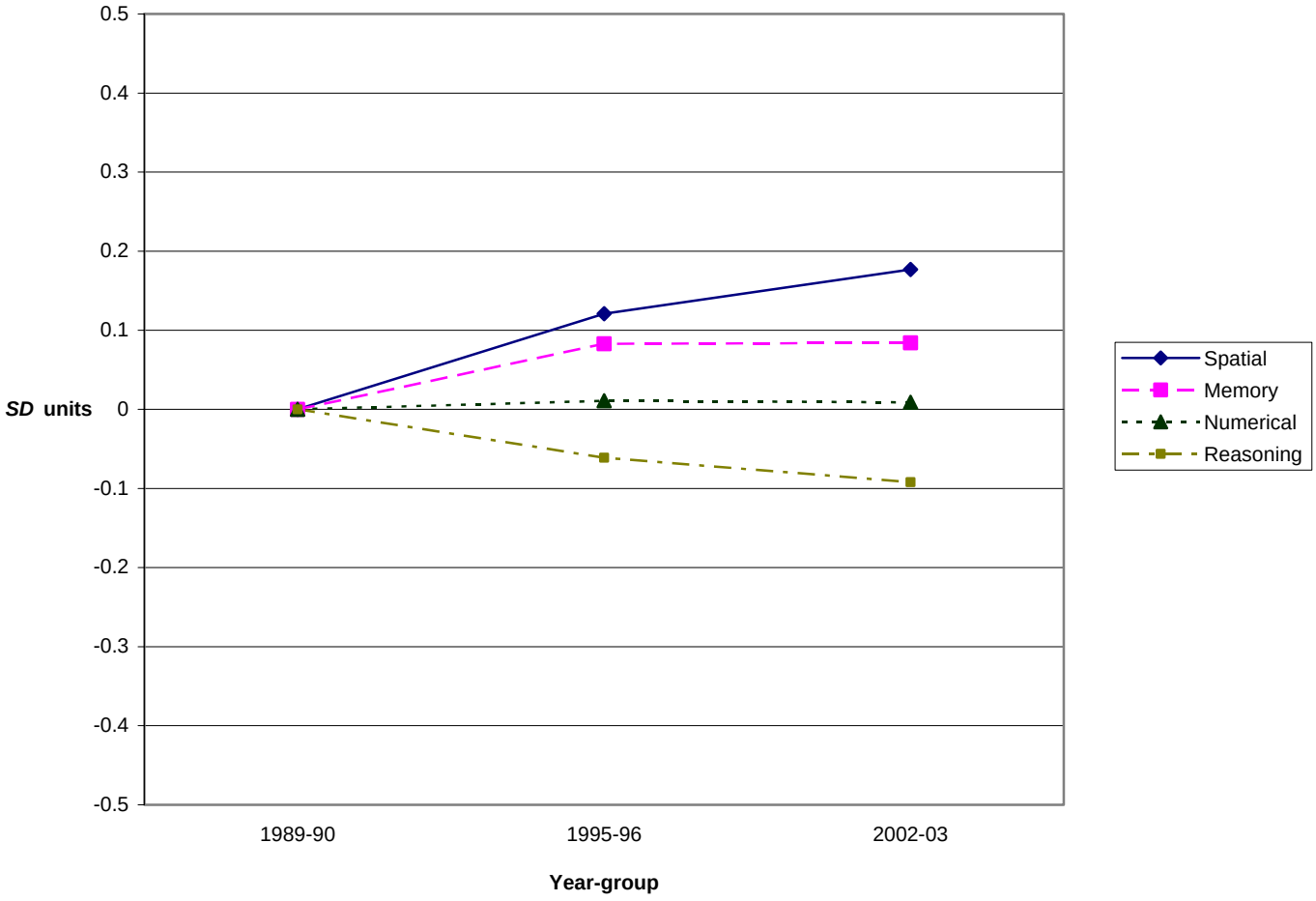


Figure 4

Changes in General Aptitude Score From 1989-90 to 2002-03 by Sex

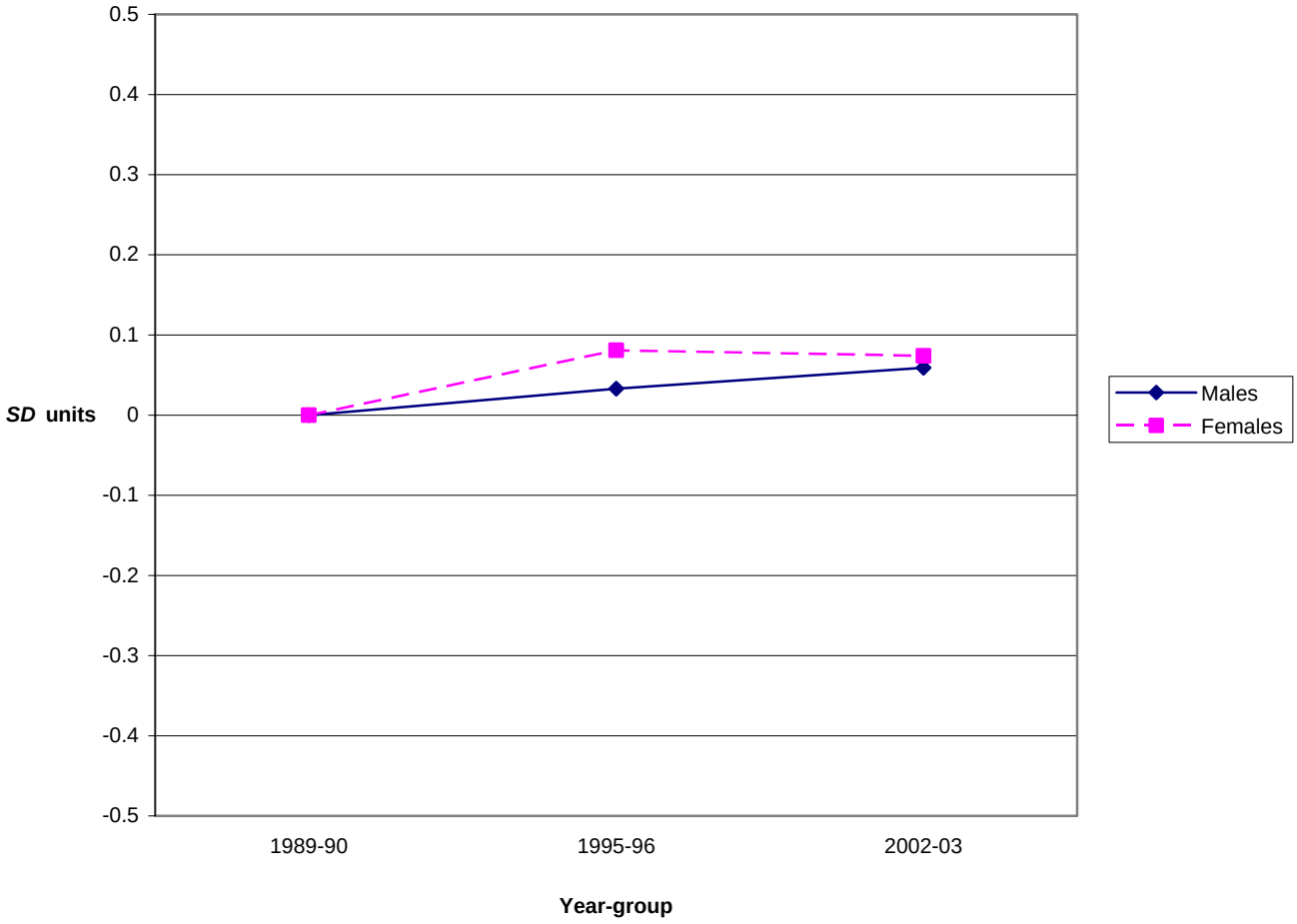


Figure 5

Changes in Domain Scores From 1989-90 to 2002-03 for Females

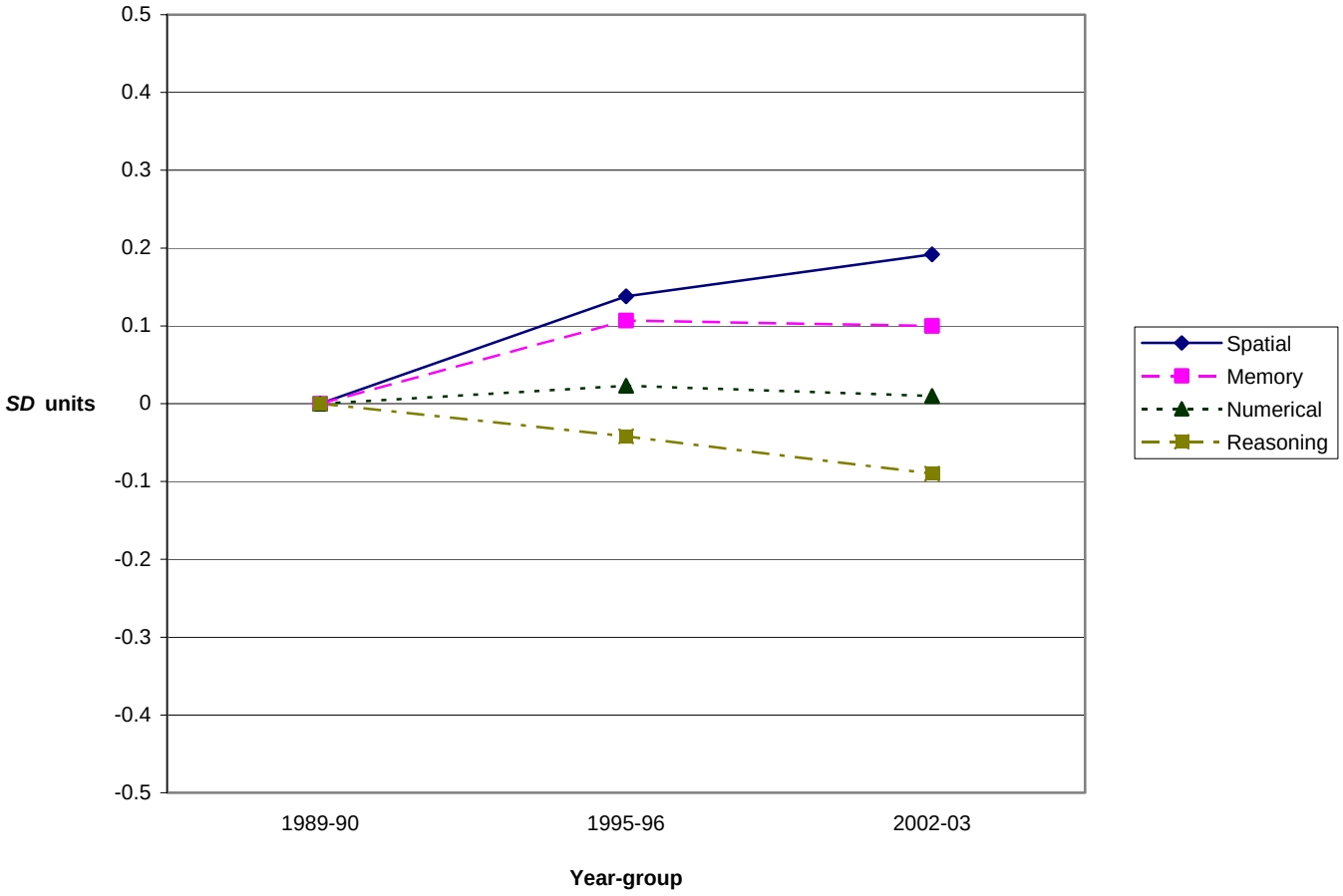


Figure 6

Changes in Domain Scores From 1989-90 to 2002-03 for Males

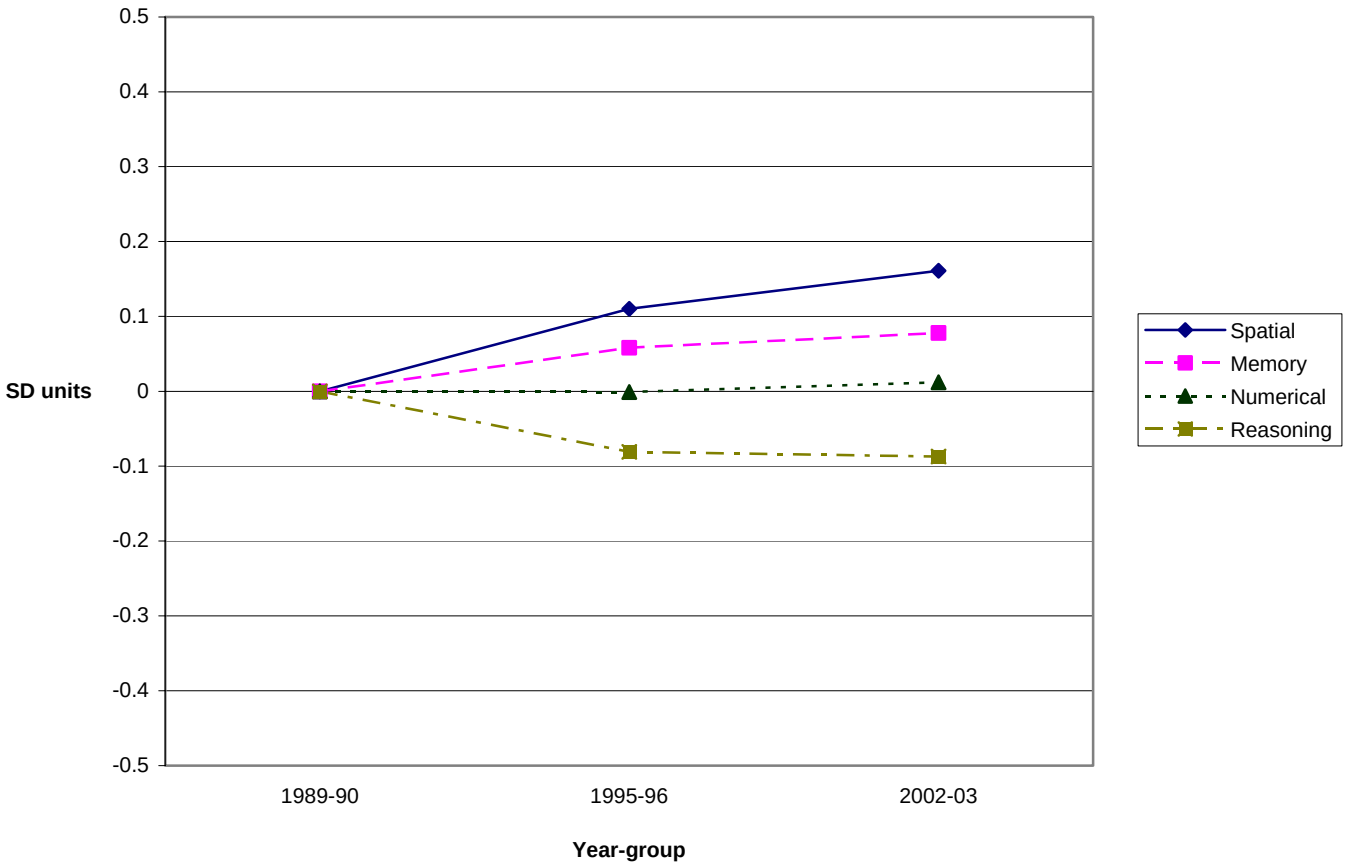
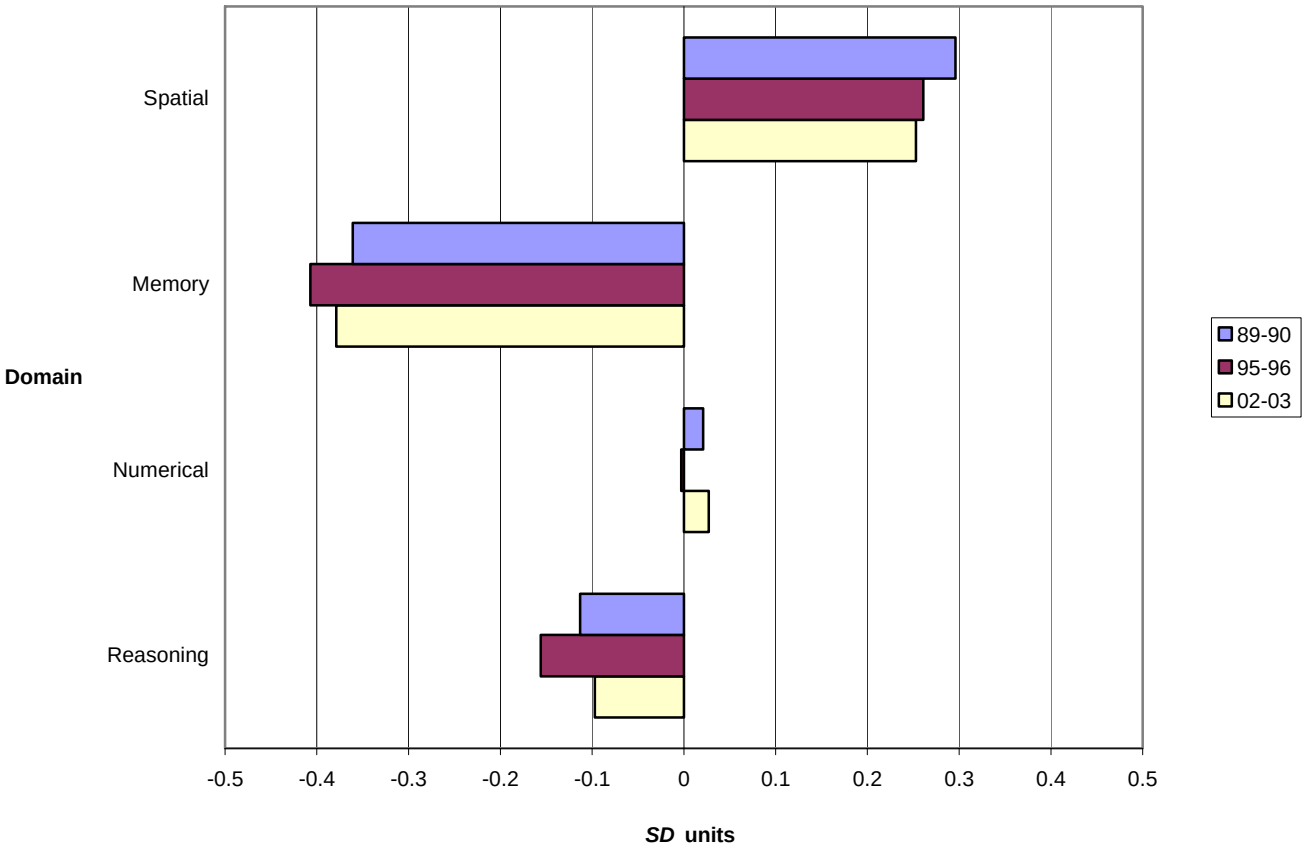


Figure 7

Differences Between Males and Females by Domains From 1989-90 to 2002-03



Note. The bars extending to the right of zero indicate higher performance by males than females; bars extending to the left indicate higher performance by females.

Figure 8

Changes in General Aptitude Score From 1989-90 to 2002-03 by Age Group

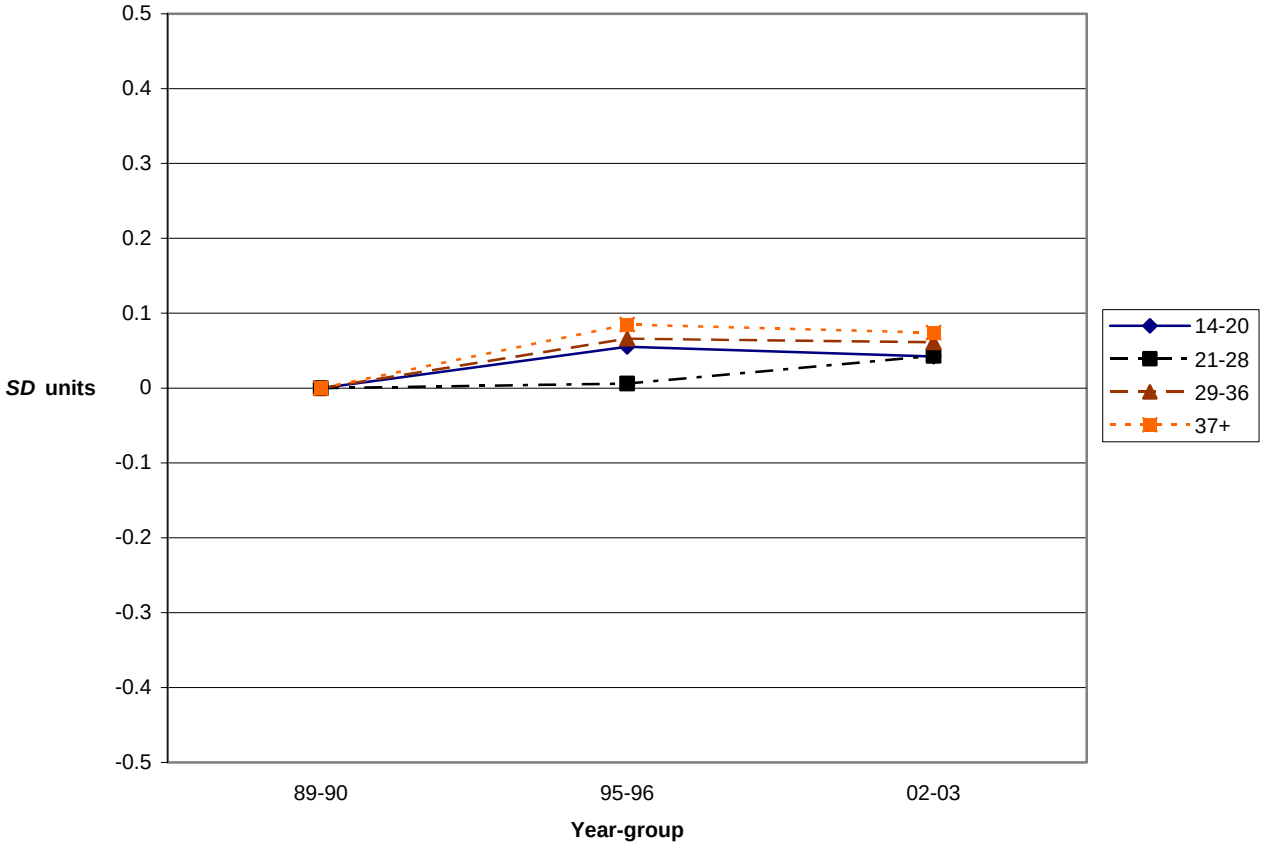


Figure 9

Changes in Spatial Domain Score From 1989-90 to 2002-03 by Age Group

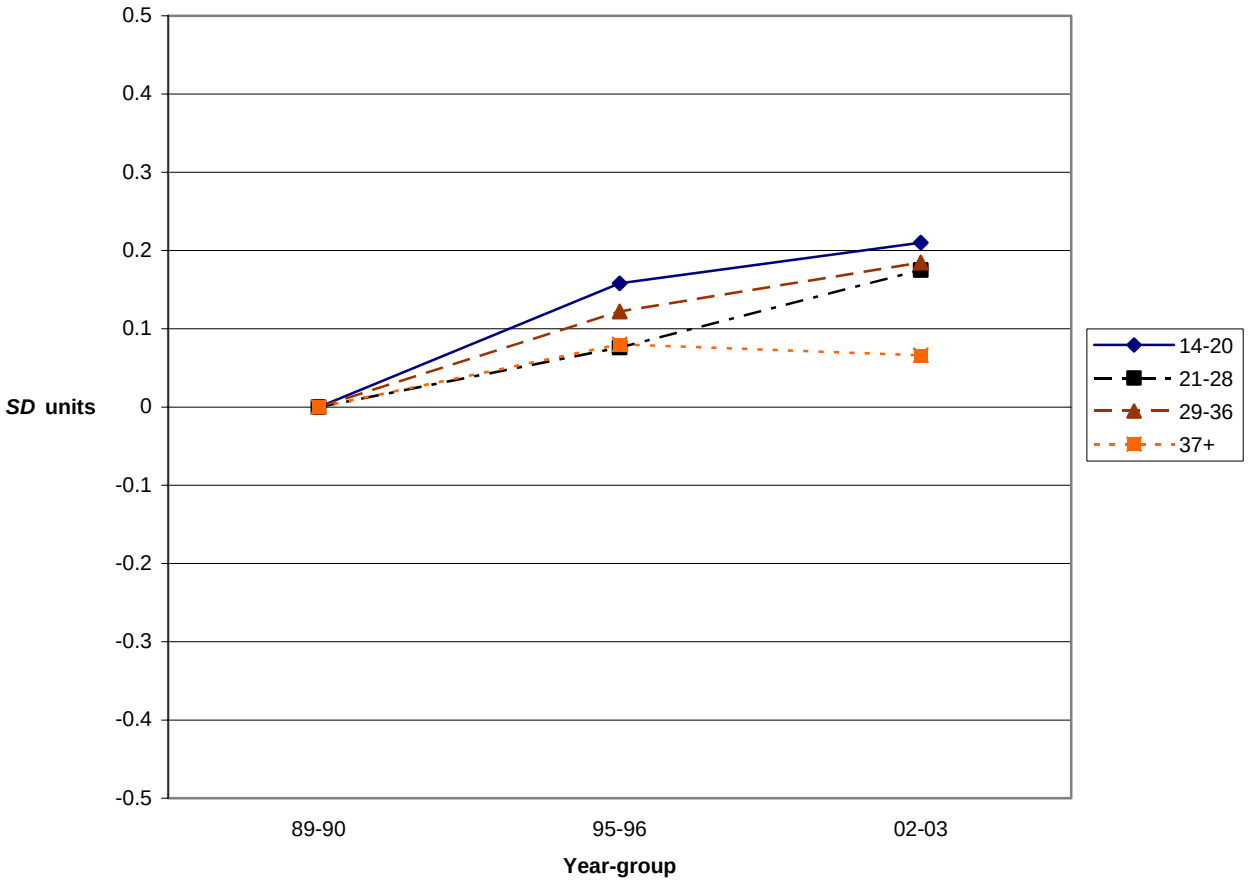
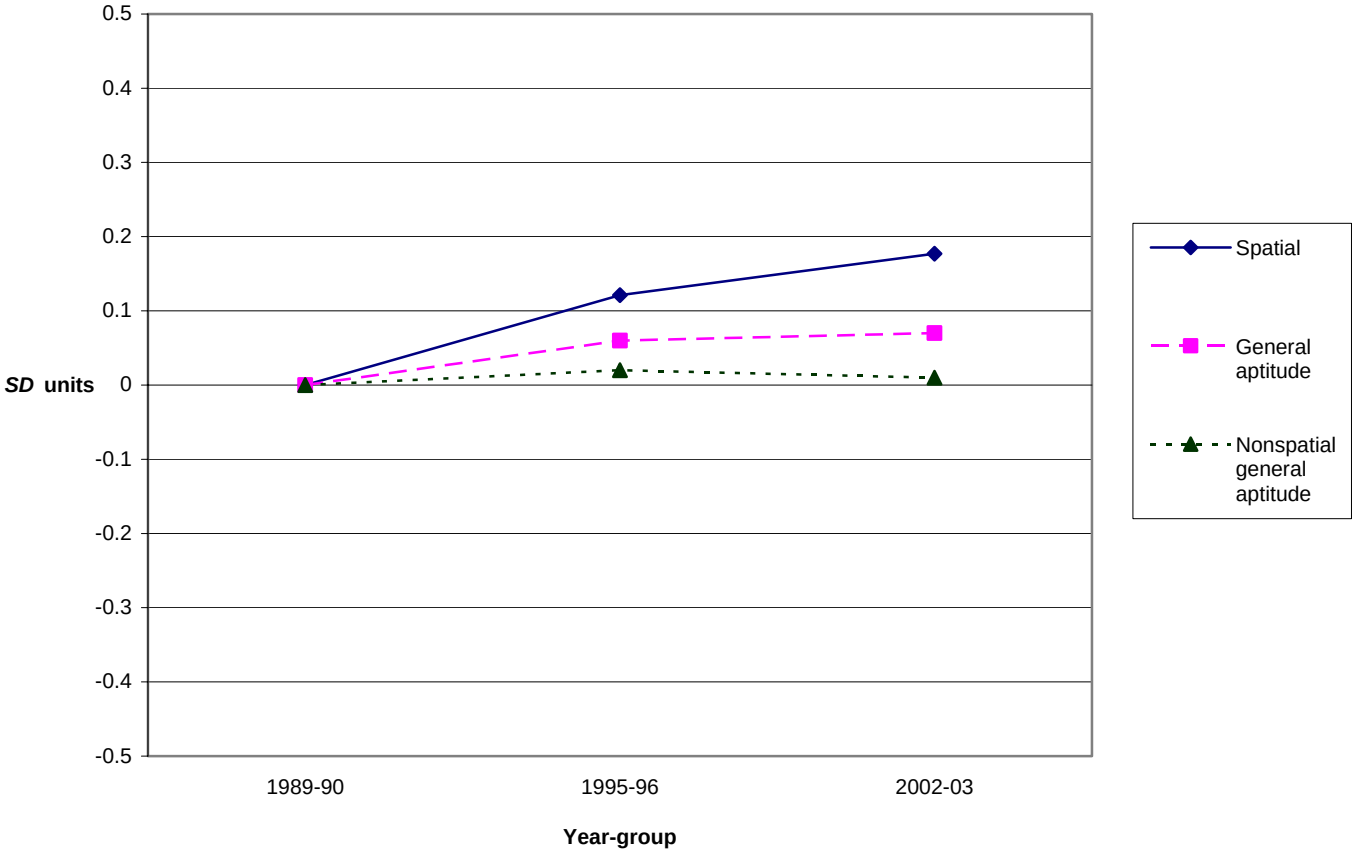


Figure 10

Changes in Mean Standardized Scores for General Versus Spatial and Nonspatial General Aptitude From 1989-90 to 2002-03



Appendix A

Changes in Mean Standardized Test Scores From 1989-90 to 2002-03

	Year-group		
Test score	1989-90	1995-96	2002-03
Spatial			
Paper Folding	.00	.08	.11
Memory for Design	.00	.13	.20
Memory			
Silograms	.00	.10	.07
Number Memory	.00	.10	.03
Observation	.00	-.01	.08
Numerical			
Number Series	.00	.09	.12
Number Facility	.00	-.07	-.11
Reasoning			
Inductive Reasoning	.00	-.09	-.17
Analytical Reasoning	.00	-.01	.02

Note. *N* for 1989-90 = 10,969, *N* for 1995-96 = 8,092, *N* for 2002-03 = 6,703. The values for 1989-90 were set to zero so that the other values represent mean changes relative to 1989-90. For 2002-03, all nine test means differed significantly from each other.

Appendix B

Changes in Mean Standardized Test Scores From 1989-90 to 2002-03 by Sex

	Sex					
	Male			Female		
Test score	89-90	95-96	02-03	89-90	95-96	02-03
Spatial						
Paper Folding	.00	.07	.08	.00	.09	.15
Memory for Des.	.00	.12	.20	.00	.15	.19
Memory						
Silograms	.00	.10	.09	.00	.11	.07
Number Memory	.00	.07	.00	.00	.12	.08
Observation	.00	-.04	.09	.00	.01	.08
Numerical						
Number Series	.00	.09	.13	.00	.09	.11
Number Facility	.00	-.05	-.11	.00	-.05	-.10
Reasoning						
Inductive Reas.	.00	-.12	-.17	.00	-.06	-.16
Analytical Reas.	.00	-.02	.03	.00	-.01	.01

Note. Male: *N* for 1989-90 = 5,754, *N* for 1995-96 = 4,215, *N* for 2002-03 = 3,595. Female: *N* for 1989-90 = 5,215, *N* for 1995-96 = 3,877, *N* for 2002-03 = 3,108. The values for 1989-90 were set to zero so that the other values represent mean changes relative to 1989-90. For 2002-03, within each sex, all nine test means differed significantly from each other.

Appendix C

Changes in Mean Standardized Test Scores From 1989-90 to 2002-03 by Age Group

	Year-group		
Test score	1989-90	1995-96	2002-03
Spatial			
Paper Folding			
14-20	.00	.12	.16
21-28	.00	.04	.11
29-36	.00	.09	.10
37+	.00	.04	.01
Memory for Design			
14-20	.00	.16	.20
21-28	.00	.09	.21
29-36	.00	.13	.23
37+	.00	.10	.11
Memory			
Silograms			
14-20	.00	.06	.04
21-28	.00	.11	.09
29-36	.00	.08	.02
37+	.00	.15	.10
Number Memory			
14-20	.00	.10	-.01
21-28	.00	.04	.02
29-36	.00	.12	.03
37+	.00	.09	.08
Observation			
14-20	.00	-.01	.06
21-28	.00	-.03	.14
29-36	.00	-.03	.09
37+	.00	.00	.04
Numerical			
Number Series			
14-20	.00	.15	.19
21-28	.00	.00	.08
29-36	.00	.06	.02
37+	.00	.11	.08

Test score	Year-group		
	1989-90	1995-96	2002-03
Number Facility			
14-20	.00	-.13	-.23
21-28	.00	-.14	-.16
29-36	.00	.00	-.02
37+	.00	.02	.11
Reasoning			
Inductive Reasoning			
14-20	.00	-.12	-.20
21-28	.00	-.13	-.22
29-36	.00	-.05	-.11
37+	.00	-.05	-.17
Analytical Reasoning			
14-20	.00	-.04	-.01
21-28	.00	-.02	.00
29-36	.00	-.03	.00
37+	.00	.03	.07

Note. *N*s for 14-20: 1989-90 = 3,787; 1995-96 = 2,947; 2002-03 = 3,174. *N*s for 21-28: 1989-90 = 3,061; 1995-96 = 2,263; 2002-03 = 1,665. *N*s for 29-36: 1989-90 = 1,840; 1995-96 = 1,211; 2002-03 = 759. *N*s for 37+: 1989-90 = 2,281; 1995-96 = 1,671; 2002-03 = 1,105. The values for 1989-90 were set to zero so that the other values represent mean changes relative to 1989-90. For 2002-03, the four means for Number Series, Number Facility, Paper Folding, Memory for Design, and Observation all differed significantly from each other.

Appendix D

Latent-Variable Analysis for Secular-Change Data

As noted in the text, in the body of this report we describe analyses of observed, or manifest, variables (that is, raw scores or transformations of raw scores), whereas in this appendix we describe analyses of latent variables (or factors). Latent-variable analysis permits one to explicitly define how the overlapping variance between variables, such as the spatial domain score and the general aptitude score, should be handled in addition to taking advantage of a number of other technical enhancements (e.g., automatic correction for unreliability). For these analyses, we collaborated with Dr. Jelte Wicherts, of the University of Amsterdam.

The factor model that we used for our nine tests is shown in the accompanying figure. The model contains a second-order general aptitude factor with four first-order factors (reasoning, numerical, spatial, memory) and 2-3 individual tests per factor. In previous research, we demonstrated that this model shows moderately good fit for the Foundation's tests (Condon & Schroeder, 2003).

As noted in the Method section, we had samples of Foundation examinees for the nine tests for each of three time periods: 1989-90, 1995-96, and 2002-03. In our analyses here, we assumed that the factor model just described could be fit to the data for each time period and that the time periods would differ only in that there would be mean gains (or losses) for the five factors between Time 1 and Time 2 and between Time 2 and Time 3. The type of analysis most appropriate for data like this is *multiple-group confirmatory factor analysis* (MGCFA), which we used. This procedure has been described by Dolan, Colom, Abad, Wicherts, Hessen, and van de Sluis (2006) and at greater length by Byrne (2001).

In the next analyses, we evaluated whether the assumptions described here were appropriate for our data. Customarily, one assesses whether the given model is suitable at the first-order level (ignoring the general factor), and if it is suitable, one subsequently evaluates the second-order level.

The table in this appendix shows the results for the fit of the first-order model. Step 1 shows the fit of the baseline model to which the other steps were compared. The fit of the baseline model is marginally good. The Comparative Fit Index (CFI) is above .90 and the Standardized Root Mean Square Residual (SRMR) is .05, but the Root Mean Square Error of Approximation (RMSEA) is .103. For reference, a good-fitting model would show a CFI above .90, and ideally above .95, and a SRMR and a RMSEA below .05. (See Byrne [2001], for an extended discussion of fit analyses.)

Steps 2 and 3 show the fit when the factor loadings and residual variances, respectively, are restricted to be invariant across the three cohorts. Both the factor loadings and residual variances appear to be invariant in this way. The problem for the

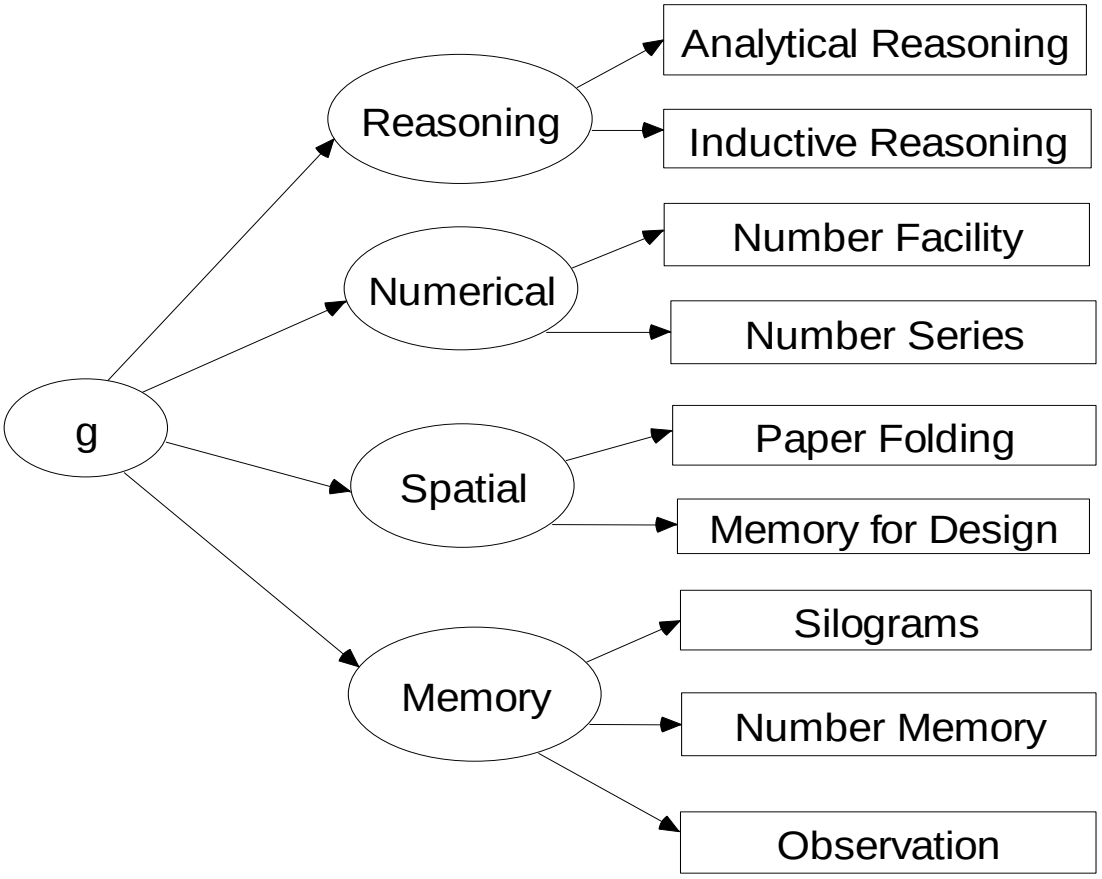
first-order model comes in Step 4, when we examined the restriction of equal intercepts across intercepts. In this step, the SRMR and CAIC increase and the CFI decreases, which indicates worse fit. The modification indexes show some reasons for the poorer fit: for example, for the memory factor, Observation shows greater gains between 1989-90 and 2002-03 than do Silograms and Number Memory (and so, as a group these tests do not maintain "equal intercepts" across the three time periods). This is not an error,[7] but rather an indication of a small misalignment between the MGCFA model that we intended to use and the structure of the test score data. In a similar way, for the reasoning factor, Inductive and Analytical Reasoning showed somewhat different patterns across time, and for the numerical factor, Number Series and Number Facility showed somewhat different patterns across time.

Within the MGCFA framework, it is acceptable to modify the model to permit different tests to have different intercepts over time, but this makes it impossible to identify substantive gains in performance. At this point we concluded that an MGCFA analysis was not feasible with our data. It is possible that there is an alternative latent-variable approach that would allow one to estimate gains at the factor level while also having unique effects at the level of individual tests.

[7] In fact, it is not surprising that Observation, which uses visual stimuli, shows greater gains than Silograms and Number Memory, which use words and numbers, respectively.

Figure

Structural Equations Model for Foundation Test Battery



Table

Analysis of Fit ("Strict Factorial Invariance") for First-Order Model

Step	Description	df	χ^2	SRMR 89-90	SRMR 95-96	SRMR 02-03	RMSEA	CFI	CAIC
1	Baseline model	63	6037.5**	.051	.052	.050	.105	.903	7119
2	Factor loadings invariant	73	6087.7**	.052	.052	.050	.098	.902	7052
3	Residual variances invariant	91	6183.3**	.053	.053	.052	.088	.901	6936
4	Equal intercepts	101	6691.0**	.053	.053	.053	.087	.893	7330

Note. SRMR = Standard Root Mean Square Residual, RMSEA = Root Mean Square Error of Approximation, CFI = Comparative Fit Index, CAIC = Consistent Akaike's Information Criterion. *N* for 1989-90 = 10,969; *N* for 1995-96 = 8,092; *N* for 2002-03 = 6,703.

** $p < 0.01$.