Available online at www.sciencedirect.com
ScienceDirect
Intelligence 36 (2008) 121–126
INTELLIGENCE

# Secular declines in cognitive test scores: A reversal of the Flynn Effect

Thomas W. Teasdale [a,*], David R. Owen [b]

[a] Department of Psychology, University of Copenhagen, Øster Farimagsgade 5, 1353 Copenhagen K, Denmark
[b] Department of Psychology, Brooklyn College, City University of New York, United States

Received 13 June 2006; received in revised form 30 January 2007; accepted 30 January 2007
Available online 2 March 2007

**Abstract**

Scores on cognitive tests have been very widely reported to have increased through the decades of the last century, a generational phenomenon termed the 'Flynn Effect' since it was most comprehensively documented by James Flynn in the 1980's. There has, however, been very little evidence concerning any continuity of the effect specifically into the present century. We here report data from a population, namely young adult males in Denmark, showing that whereas there were modest increases between 1988 and 1998 in scores on a battery of four cognitive tests–these constituting a diminishing continuation of a trend documented back to the late 1950's–scores on all four tests declined between 1998 and 2003/2004. For two of the tests, levels fell to below those of 1988. Across all tests, the decrease in the 5/6 year period corresponds to approximately 1.5 IQ points, very close to the net gain between 1988 and 1998. The declines between 1998 and 2003/4 appeared amongst both men pursuing higher academic education and those not doing so.


*Keywords:* Cognitive tests; Secular trend; Flynn Effect

More than 30 years ago Flynn systematically documented evidence, based on test norms of the Stanford–Binet and Wechsler tests, implying that scores on IQ tests had been rising in the United States, through the decades of the last century (Flynn, 1984). Direct evidence came from a review of population studies, in numerous developed countries, where the same test had been used on populations at different times, typically separated by many years, and showing that later generations performed better on tests than earlier generations (Flynn, 1987). The strongest evidence in this category came from countries in which draftees for military service had been tested.

The 'Flynn Effect', as it became known, has subsequently been shown to be ubiquitous and there has been much discussion about its causes, with theories ranging from those emphasising social and education changes to those emphasising more biological factors, e.g., health care and nutrition (Neisser, 1998). A recent trend has been the reporting of a Flynn Effect–from direct or indirect evidence–in developing countries (Cocodia et al., 2003; Daley, Whaley, Sigman, Espinosa, & Neumann, 2003; Meisenberg, Lawless, Lambert, & Newton, 2005; Howard, 2005).

Despite the numerous reports, several limitations are also widespread in the relevant literature. A first limitation, as pointed out by Raven (2000), is that many studies involve simply the comparison of mean scores, often using

* Corresponding author.
*E-mail address:* tom.teasdale@psy.ku.dk (T.W. Teasdale).

0160-2896/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.intell.2007.01.007

scaled and IQ scores such as those calculated in the Wechsler Intelligence tests. This may conceal differential changes at different levels of ability. There is no prima facie reason to suppose that the Flynn Effect has involved a uniform shift in mean performance with no change in variance or higher order distributional characteristics such as skewness, and there is some evidence to suggest that gains have not been uniform across all ability levels (Colom, Lluis-Font, & Andres-Pueyo, 2005). A second limitation is that most studies have compared performances on the same tests taken by samples of subjects representative of generations separated by a number of years, typically decades. The limitation of this method is that it is not possible to determine the time course of the effect; a difference between test performance, say, in the 1970's and in the 1990's, could not unambiguously be attributed to a simple linear increase over the two decades. From this follows a third limitation. Even the few studies which have reported data from the present decade, compared to some previous decade, e.g., cannot be taken as evidencing a continuing Flynn Effect up to the present day.

Using data derived from the Danish draft board we have previously reported on substantial gains in test scores, particularly at the low end of the distribution, through the 1960's and 1970's (Teasdale & Owen, 1987, 1989), that by the 1990's they had substantially plateaued (Teasdale & Owen, 2000) and indeed have begun to decline somewhat into the present decade (Teasdale & Owen, 2005). A simultaneous plateau had been reported for Norwegian conscripts by Sundet, Barlaug, and Torjussen (2004) for tests of a range of cognitive abilities. Our own prior report of the present plateau had concerned compound scores on a draft board test summed over four subtests of differing cognitive functions which could be characterised as logical, verbal, numerical and spatial reasoning. The lack of change in the compound score into the present decade could, however, have masked different and possibly opposite trends in these separate abilities. Wicherts et al. (2004) have emphasised the importance of examining Flynn Effect evidence at the level of subtests. The primary objective of the present study has therefore been to examine the recent secular trends in each of these abilities considered separately. A secondary objective has been to explore the relationship between recent changes and educational level.

## 1. Method

Our data stem from the records of the Danish draft board. There has been conscription in Denmark continuously since the Second World War, and on attaining the age of 18 or shortly thereafter young men are required to appear before a draft board which assesses their suitability for military service. About 5–10% are exempted from appearing in person, these being largely men who can document a disqualifying illness, e.g., asthma and Scheuermann's disease.

Ever since 1957, and continuing to the present day, the draft board assessment has included an unchanged set of four group-administered tests of cognitive abilities, collectively termed Børge Prien's Prøve. The first test, Letter Matrices (19 items, 15 min), resembles Raven's Progressive Matrices with the important difference that patterns of alphabetic letters are used and the correct answer is to be supplied by the testee, rather than chosen from a set of forced-choice alternatives. A Verbal Analogies test (24 items, 5 min) comprises a series of analogies somewhat akin to Miller's Analogies test but where the correct response is to be found in an alphabetically arranged list of 100 words. In a Number Series test (17 items, 15 min), the fifth number following a series of four numbers is to be deduced, and in a Geometric Figures test (18 items, 10 min) a set of complex geometric shapes are to be partitioned into simpler components. All of the tests are scored as the number of correct responses and a total score (0 through 78) is also calculated; this total has been found to correlate 0.8 with the Wechsler Adult Intelligence Scale (Mortensen, Reinisch, & Teasdale, 1989). Further details of the tests are presented elsewhere (Rasch, 1980; Teasdale & Owen, 1989).

Also recorded by the draft board is level of school education. The coding for this was changed substantially in 1991 and the 1988 cohort cannot therefore be compared in this respect with the later two cohorts. The coding is also sensitive to seasonal variation since men tested in the second half of the year are more likely to have completed their ultimate level of schooling. We have therefore here employed a dichotomized index of level of schooling which is independent of season, namely whether or not the subject had attended a 'Gymnasium' (or some equivalent), i.e., an advanced school for 16–18 year-olds leading towards a university entrance and other forms of higher education. Students not attending a Gymnasium typically enter at age 15 or 16 and have shorter and non-academic occupational training courses.

In this report we present data on all men who were tested in 1988 ($n$=33,833), 1998 ($n$=25,020) and in the second half of 2003 together with the first half of 2004, here designated 2003/4 ($n$=23,598) respectively. The difference among the $n$s in these three cohorts is predominantly due to the declining birth-rate in

Table 1
Standardized test scores in relation to year of testing

| | Year Tested | Mean | Standard deviation | Skewness | Kurtosis |
|---|---|---|---|---|---|
| Letter Matrices | 1988 | 0.000 | 1.000 | −0.634 | 0.243 |
| | 1998 | 0.076 | 0.951 | −0.671 | 0.414 |
| | 2003–4 | 0.032 | 0.981 | −0.632 | 0.291 |
| Verbal Analogies | 1988 | 0.000 | 1.000 | −0.249 | 0.095 |
| | 1998 | 0.065 | 0.978 | −0.350 | 0.232 |
| | 2003–4 | −0.016 | 0.982 | −0.313 | 0.110 |
| Number Series | 1988 | 0.000 | 1.000 | −0.560 | −0.272 |
| | 1998 | 0.059 | 0.976 | −0.637 | −0.118 |
| | 2003–4 | −0.059 | 0.990 | −0.509 | −0.348 |
| Geometric Figures | 1988 | 0.000 | 1.000 | 0.034 | 0.166 |
| | 1998 | 0.160 | 1.000 | 0.048 | 0.033 |
| | 2003–4 | 0.087 | 0.998 | 0.049 | 0.111 |
| IQ | 1988 | 100.000 | 15.000 | −0.441 | 0.134 |
| | 1998 | 101.650 | 14.559 | −0.525 | 0.336 |
| | 2003–4 | 100.160 | 14.725 | −0.490 | 0.234 |

Denmark, although in some part it is also due to increasing proportions whose medical unsuitability is registered by the draft board without their being required to appear personally.

In order to facilitate comparisons we have transformed the raw scores from the four tests into *z*-scores standardized against the 1988 means and standard deviations. Similarly, we have transformed the total scores into deviation IQs, standardizing on the 1988 data, setting that mean to 100 and standard deviation to 15.

## 2. Results

Table 1 shows the mean test scores and IQs for the 1988, 1998 and 2003/4 cohorts. All four tests and IQ increased significantly ($p<.001$) between 1988 and 1998. The magnitude of the ten-year gains, however, was modest, being less than 0.1 standard deviations for three of the four tests, and about 0.16 standard deviations for the Geometric Figures test. Overall IQ gain over the ten-year time-period was less than two points.

Between 1998 and 2003/4 mean scores on all four tests and IQ declined significantly ($p<.001$). The smallest decline was found for the Geometric Figures test whereas performance on the Verbal Analogies and Number Series tests actually declined to below the 1988 level. In the years between 1998 and 2003/4, mean IQ fell significantly ($p<.001$) by about 1.5 points, corresponding to almost all of the gain in the decade 1988 to 1998.

During the 1960's and decades up to the 1990's we have observed declining variances and increasing negative skewness. Examination of low and high percentiles revealed that the test score gains was primarily occurring at the low end of the distributions with gradually fewer and fewer low scores. As can be seen this trend continued between 1988 and 1998 for three of the tests, namely the Logical Matrices, Verbal Analogies and Number Series tests, and for IQ itself. Between 1998 and 2003/4 these trends have been reversed. The Geometric Figures test was exceptional in this respect also.

It is worth noting that none of the four tests have suffered any manifest ceiling effect. For all four tests, at their height in 1998, it is the case that fewer than 10% of the men tested correctly answered more than at most 83% of the items.

Table 2 shows the intercorrelations (Pearson's Product-Moment and Point-Biserial, as appropriate) among the four tests. There are statistically significant differences among the three $\Sigma$ matrices (Bartlett's test for homogeneity of variance–covariance matrices, $\chi^2(20)=155$, $p<.001$). This in part reflects the distributional changes across time over the four tests and in

Table 2
Test intercorrelations in relation to year of testing

| | Letter Matrices | Verbal Analogies | Number Series | Geometric Figures | IQ |
|---|---|---|---|---|---|
| *1988* | | | | | |
| Verbal Analogies | .57 | | | | |
| Number Series | .61 | .61 | | | |
| Geometric Figures | .48 | .49 | .45 | | |
| IQ | .80 | .86 | .82 | .75 | |
| *1998* | | | | | |
| Verbal Analogies | .56 | | | | |
| Number Series | .61 | .59 | | | |
| Geometric Figures | .47 | .47 | .43 | | |
| IQ | .79 | .85 | .81 | .74 | |
| Gymnasium attendance | .43 | .50 | .47 | .29 | .53 |
| *2003/4* | | | | | |
| Verbal Analogies | .58 | | | | |
| Number Series | .62 | .59 | | | |
| Geometric Figures | .47 | .46 | .43 | | |
| IQ | .80 | .85 | .82 | .74 | |
| Gymnasium attendance | .42 | .48 | .47 | .26 | .51 |

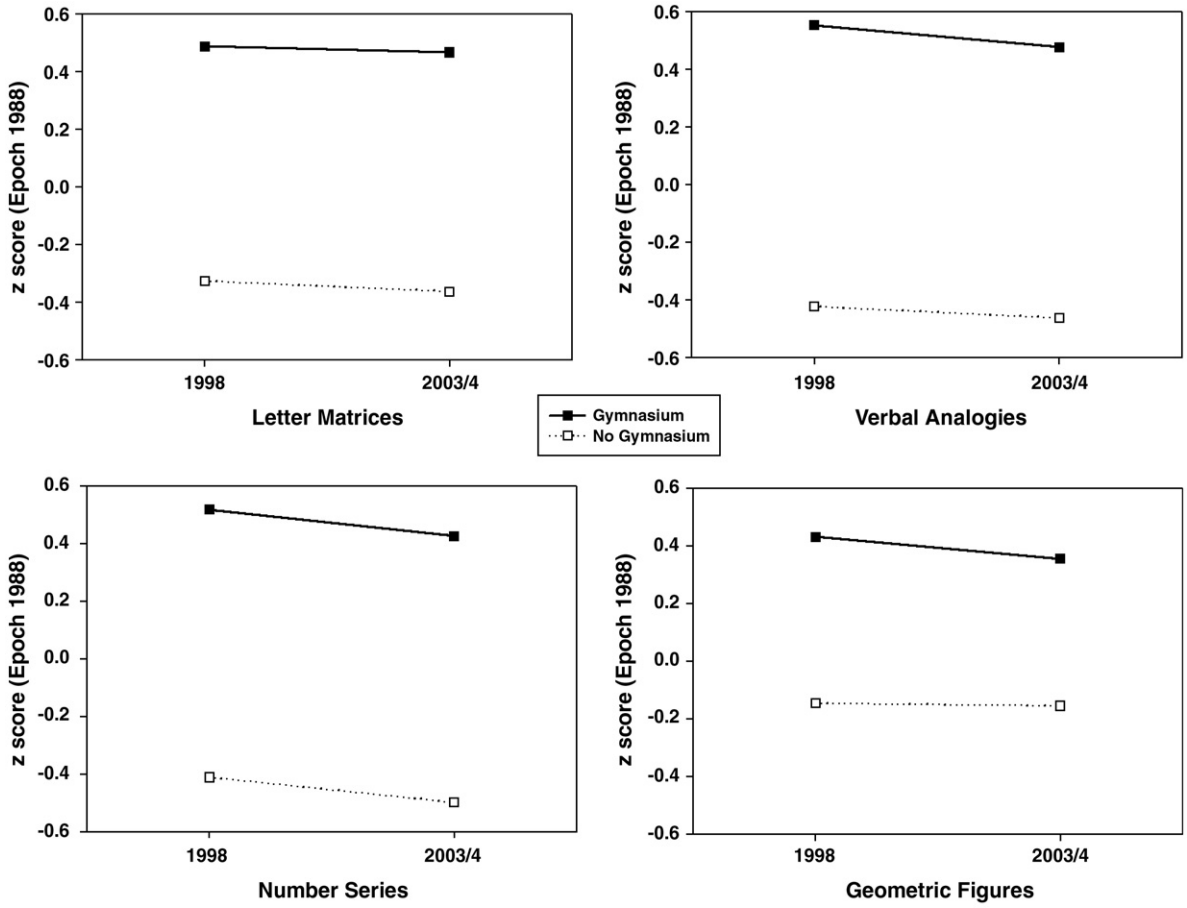


Fig. 1.

part the considerable statistical power derived from the large number of observations involved. Primarily, however, it can be seen that the correlations have been markedly stable across the 15-year time-period. Furthermore, for the four tests excluding IQ, the first maximum-likelihood factor accounts for 64–65% of the variance within each of the three cohorts separately, indicating a uniformly very substantial *g* loading.

We have reported elsewhere that 1998 represented a high point in Gymnasium attendance (48.7%), and immediately a decline in the following years had suggested the possibility that the simultaneous decline in test scores had been a consequence of this (Teasdale & Owen, 2005). However, Gymnasium attendance in 2003/4 had risen again to almost the same level (47.7%), making this conjecture now less likely. In fact, as shown in Fig. 1, the subtest score declines since 1998 have been present both for men who had attended Gymnasium and those who had not done so. These declines were significant or near significant ($p < .055$) in all cases with the single exception of the comparison for non-gymnasium attendees on the Geometric Figures test.

## 3. Discussion

The major conclusion from the present study is a confirmation that, after several decades of increasing cognitive test performance consistent with the apparently ubiquitous Flynn Effect, Danish test scores have declined, albeit modestly, within the first years of the present decade. This finding concurs with that from another Scandinavian country, namely Norway (Sundet et al., 2004) with which Denmark shares many historical, linguistic, cultural and social characteristics.

The declines across the four tests do not seem to be the results of artifactual influences. Because of the abstract nature of the items in each test, it is unlikely that obsolescence is making them more difficult for present generations that for earlier ones. Similarly, none of the four tests appears to have suffered from any clear ceiling effect, as for instance has been the case with Raven's

(2000) Progressive Matrices, and even had that been the case it would not account for an actual decline in tests scores. It is sometimes argued that draft board tests, because of the circumstances under which they are administered, are susceptible to malingering, and the incidence of this might vary across time. In an unreported study, however, we have found that an expressed negative attitude to performing military conscription was actually associated with somewhat higher rather than lower test scores. This apparently paradoxical association could be fully accounted for statistically by the fact that young men who had attended gymnasiums, and who, as shown, score much higher on the tests that those who do not, were less likely to be positively motivated to perform conscription since it would interrupt any progress to further education such as universities.

Although the declines therefore seem to be real, it is not easy to account for them, particularly in view of the fact that broad unanimity has never been reached on the causes of the Flynn Effect itself. Improving nutrition has sometimes been proposed as a factor in the effect, particularly in developing countries (Arija et al., 2006; Colom et al., 2005; Daley et al., 2003), but it does not seem plausible that there has been any decline in nutrition in Denmark over the past two decades; certainly the mean height of conscripts, a good index of nutrition, has shown no such decline and has in fact remained very stable at about 180 cm since 1985 (Larnkjaer, Schrøder, Schmidt, Jørgensen, & Michaelsen, 2006).

We consider that the Flynn Effect, at least as it occurred in Denmark, was primarily to be attributed to social and, in particular, educational improvements, including resources allocated to special education. Such a contention is certainly prima facie supported by the strong associations invariably found between cognitive test performance and educational level (Barber, 2005; Blair, Gamson, Thorne, & Baker, 2005) as also seen in the present study. These associations make it, in our view, likely that the Flynn Effect, when and where it operates, reflects to an important extent genuine gains in cognitive abilities, although some test-specific factors are undoubtedly also involved. It is possible that the small secular decline in test scores found here has resulted from some qualitative changes in the emphasis on abstract reasoning and problem-solving within the Danish educational system or a decreased emphasis on speed.

Another partial contributing factor to the recent decline could be the ethnic composition of young Danes, specifically the rising proportion who are immigrants or their immediate descendants. te Nijenhuis, de Jong, Evers, and van der Flier (2004) have reviewed data from the Netherlands, showing that children of immigrants do not generally perform as well on cognitive tests as children who are ethnically Dutch, although they do show improvements over first generation immigrants. A recent unpublished study of Danish draftees has similarly shown that immigrant groups (first or second generation with Danish nationality) score below overall averages on all of the four tests reported on here. The proportion of such immigrants among the Danish male 18-year-old population rose from less than 1% in 1988 to about 2.3% in 1998 and 5% in 2004 (http://www.statistikbanken.dk).

It is important to recognise, however, that, if indeed the Flynn Effect is now at an end in such highly developed countries as Norway and Denmark, it may be far from over in countries which are less developed. Much of the recent reporting of a continued Flynn Effect has come from such countries (Cocodia et al., 2003; Daley et al., 2003; Meisenberg et al., 2005). If such differentials were to continue, then any national differences in cognitive test performances (Lynn & Vanhanen, 2002) might be expected to diminish in the future.

## References


Arija, V., Esparo, G., Fernandez-Ballart, J., Murphy, M. M., Biarnes, E., & Canals, J. (2006). Nutritional status and performance in test of verbal and non-verbal intelligence in 6 year old children. *Intelligence*, *34*, 141–149.

Barber, N. (2005). Educational and ecological correlates of IQ: A cross-national investigation. *Intelligence*, *33*, 273–284.

Blair, C., Gamson, D., Thorne, S., & Baker, D. (2005). Rising mean IQ: Cognitive demand of mathematics education for young children, population exposure to formal schooling, and the neurobiology of the prefrontal cortex. *Intelligence*, *33*, 93–106.

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.

Colom, R., Lluis-Font, J. M., & Andres-Pueyo, A. (2005). The generational intelligence gains are caused by decreasing variance in the lower half of the distribution: Supporting evidence for the nutrition hypothesis. *Intelligence*, *33*, 83–91.

Daley, T. C., Whaley, S. E., Sigman, M. D., Espinosa, M. P., & Neumann, C. (2003). IQ on the rise— The Flynn effect in rural Kenyan children. *Psychological Science*, *14*, 215–219.

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.

Howard, R. W. (2005). Objective evidence of rising population ability: A detailed examination of longitudinal chess data. *Personality and Individual Differences*, *38*, 347–363.

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

Larnkjaer, A., Schrøder, S. A., Schmidt, I. A., Jørgensen, M. H., & Michaelsen, K. F. (2006). Secular change in adult stature has come to a halt in northern Europe and Italy. *Acta Paediatrica*, *95*, 754−755.

Meisenberg, G., Lawless, E., Lambert, E., & Newton, A. (2005). The Flynn effect in the Caribbean: Generational change of cognitive test performance in Dominica. *Mankind Quarterly*, *46*, 29−69.

Mortensen, E. L., Reinisch, J. M., & Teasdale, T. W. (1989). Intelligence As Measured by the Wais and A Military Draft Board Group Test. *Scandinavian Journal of Psychology*, *30*, 315−318.

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

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*, 2nd ed. Chicago: University of Chicago Press.

Raven, J. (2000). The Raven's Progressive Matrices: Change and stability over culture and time. *Cognitive Psychology*, *41*, 1−48.

Sundet, J. M., Barlaug, D. G., & Torjussen, T. M. (2004). The end of the Flynn effect? A study of secular trends in mean intelligence test scores of Norwegian conscripts during half a century. *Intelligence*, *32*, 349−362.

Teasdale, T. W., & Owen, D. R. (1987). National secular trends in intelligence and education— A 20-year cross-sectional study. *Nature*, *325*, 119−121.

Teasdale, T. W., & Owen, D. R. (1989). Continuing secular increases in intelligence and a stable prevalence of high intelligence levels. *Intelligence*, *13*, 255−262.

Teasdale, T. W., & Owen, D. R. (2000). Forty-year secular trends in cognitive abilities. *Intelligence*, *28*, 115−120.

Teasdale, T. W., & Owen, D. R. (2005). A long-term rise and recent decline in intelligence test performance: The Flynn Effect in reverse. *Personality and Individual Differences*, *39*, 837−843.

te Nijenhuis, J., de Jong, M. -J., Evers, A., & van der Flier, H. (2004). Are cognitive differences between immigrant and majority groups diminishing? *European Journal of Personality*, *18*, 405−434.

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