Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
Intelligence 32 (2004) 349–362
INTELLIGENCE

The end of the Flynn effect? A study of secular trends in mean intelligence test scores of Norwegian conscripts during half a century

Jon Martin Sundet[a,*], Dag G. Barlaug[b], Tore M. Torjussen[b]

[a]*Institute of Psychology, University of Oslo, P.O. Box 1094, Blindern N-317 Oslo, Norway*
[b]*Psychological Services, Norwegian Armed Forces, Norway*

Received 22 October 2003; received in revised form 5 April 2004; accepted 8 June 2004

Abstract

The present paper reports secular trends in the mean scores of a language, mathematics, and a Raven-like test together with a combined general ability (GA) score among Norwegian (male) conscripts tested from the mid 1950s to 2002 (birth cohorts ≈1935–1984). Secular gains in standing height (indicating improved nutrition and health care) were also investigated. Substantial gains in GA were apparent from the mid 1950s (test years) to the end 1960s–early 1970s, followed by a decreasing gain rate and a complete stop from the mid 1990s. The gains seemed to be mainly caused by decreasing prevalence of low scorers. From the early 1970s, the secular gains in GA were almost exclusively driven by gains on the Raven-like test. However, even the means on this particular test stopped to increase after the mid to late 1990s. It is concluded that the Flynn effect may have come to an end in Norway. Height gains were strongly correlated with intelligence gains until the cessation of height gains in the conscript cohorts towards the end of the 1980s. Contrary to the intelligence gains, the height gains (conscript cohorts 1969–2002) were most pronounced in the upper half of the distribution. Evidence indicating decreasing intercorrelations between tests is reported.

Keywords: Flynn effect; Intelligence; Norwegian conscripts

* Corresponding author.
E-mail address: j.m.sundet@psykologi.uio.no (J.M. Sundet).

0160-2896/$ - see front matter
doi:10.1016/j.intell.2004.06.004

1. Introduction

It has long been known among intelligence test users that test performance improves from one generation to the next (the Flynn effect), necessitating new and stricter norms from time to time. Scientific interest in secular increases of intelligence test scores virtually exploded after the publication of the seminal paper by Flynn (1987), reviewing data showing substantial gains in 14 industrialized countries in Europe, North America, and the Far East. Later, secular gains have been observed in Sweden (Emanuelsson, Reuterberg, & Svensson, 1993; Emanuelsson & Svensson, 1990), Denmark (Teasdale & Owen, 1989), Israel (Flynn, 1998a), and in urban regions in Brazil and China (Flynn, 1998b). The average gain seems to be about 3–5 IQ points per decade. Recently, a diminishing growth rate in the birth cohorts 1940–1980 of male conscripts has been observed in Denmark (Teasdale & Owen, 2000). In Sweden, the secular trends may have more or less leveled out in the birth cohorts between 1972 and 1977 (Emanuelsson et al., 1993).

The secular gains seem to be largest on tests not clearly related to school curricula and presumably measuring fluid intelligence (Cattell, 1987). On Raven's Progressive Matrices and Raven-like tests, gains in the neighborhood of 18–20 IQ points in a generation seem to be quite typical in many industrialized countries (Flynn, 1999). Kenyan 7-year-old school children in a rural district showed the most dramatic gains on the Raven Progressive Matrices Test ever observed (estimated to be at least 0.8 IQ points per year) over a 14-year period from 1984 to 1998 (Daley, Whaley, Sigman, Espinosa, & Neumann, 2003).

At least in some countries, the secular gains seem to have been unevenly distributed over ability levels. Teasdale and Owen (1989, 2000) found that the secular gains mainly were caused by lower prevalence of low scorers. In Britain, the same tendency was found for some tests (including Raven's Progressive Matrices), but not for others (Lynn & Hampson, 1986). In other countries, the gains seem to be evenly distributed cross ability levels (Flynn, 1998b).

Detterman and Daniels (1989) reported smaller correlations between different tests in high-scoring than in low-scoring groups, indicating that the secular gains may be accompanied by the changing factor structure of intelligence test scores. Recently, Kane and Oakland (2000) reported that the test intercorrelations in the U.S. Wechsler tests were lower in more recent standardization samples relative to older ones (time span 20–50 years). Indications of declining intercorrelations have also been found in France (Lynn & Cooper, 1993). Data on Danish conscripts (males) over a 10-year period showed only small downward changes in the intercorrelation pattern (Teasdale & Owen, 2000).

The main aim of the present paper is to report data on the secular trends of mean general intelligence test scores and subtest scores of a large number of Norwegian (male) conscripts who were tested in the years from 1954 to 2002, inclusive (birth cohorts $\approx$1935–1984). Changes in the distribution of test scores have also been scrutinized. In particular, we have looked for possible secular trends in the variation pattern of test scores, and, whether the secular gains have been evenly distributed across ability levels. To investigate possible changes in the prominence of the *g* factor, we have studied secular trends in the intercorrelations between tests.

In addition, secular trends of standing height have been analyzed. Standing height is a useful indicator of nutrition and health status, and secular gains in height have, to some extent, occurred in tandem with IQ gains (Martorell, 1998). Lynn (1990) has proposed that nutrition and health care improvements are among the main causal factors of IQ gains. It seems clear that nutrition and health care factors may not tell the whole story. Thus, the Flynn effect has outlasted height gains by a decade or so in many countries

(Martorell, 1998). Nevertheless, the nutrition–health care hypothesis deserves further attention. Other potential contributing factors, like more complex societies, increasing access to mass media, computer games, smaller families, changing rearing styles, to name some of the more prominent proposals, have been extensively discussed elsewhere (e.g., Neisser, 1998) and will not be further elaborated in the present paper.

2. Methods and materials

In Norway, military service is compulsory for every able young man. Before they actually enter the service, the young men are required to meet before a draft board, where their medical and psychological suitability, including intellectual ability, for military service is assessed. A great majority of the men meeting before the draft board (about 95%) are examined between their 18th and 20th birthday. The physically or psychologically disabled are exempted from these investigations. In addition, seamen and others being abroad at the normal conscript age are normally exempted.

2.1. Test materials

The draft board assessment of intellectual ability includes three speeded tests: Arithmetic (25 min), Word Similarities (8 min), and Figures (20 min). The Arithmetic test (30 items), presented in prose, purports to measure not only arithmetic and elementary algebraic ability but also logical reasoning ability, and is quite similar to the Arithmetic test in WAIS. The contents of the Arithmetic test were slightly modified in 1963. This change was mainly a modernization of some of the items, but the difficulty seems to be about the same. In the mid 1990s, the Arithmetic test was changed from open-ended answers to multiple choice (five alternatives). The Word Similarities Test (akin to the Vocabulary Test in WAIS) is a multiple-choice test (54 items). A key word is given, and the task is to find the synonym among six alternatives. The Figures Test (36 multiple choice items with six or eight alternatives) was constructed to be very similar with the Raven Progressive Matrices, except that the Raven test is organized in groups, with increasingly difficult items within each group, whereas the items in the Figures Test was constructed to provide a linear increase in difficulty (which is also the case with regard to the other tests). The Word Similarities and Figure tests have remained unchanged since 1954. The test–retest reliabilities of Arithmetic, Figures, and Word Similarities tests as calculated from a sample ($N \approx 800$) in the mid 1950s were .84, .72, and .90, respectively (Sundet, Tambs, Magnus, & Berg, 1988). The alpha coefficients of Arithmetic, Figures, and Word Similarities tests calculated for the draft cohorts 1993–2002 were .81, .80, and .90, respectively. A general ability (GA) score is a combined measure of the performance on the three tests seen together, obtained by transforming the raw scores in a standardization sample into normally distributed F scores (M=50, S.D.=20). The F scores are added and subsequently transformed into stanine scores. In a small sample (N=48), the correlation between GA and the WAIS IQ has been found to be .73 (cf. Sundet et al., 1988).

2.2. Participants and data

We have retrieved intelligence test scores from several data sets. One data set comprised GA scores for the draft cohorts from 1969 through 2001 (GA data for 1957–1959 and 2002 have been calculated from raw scores in the other data set), but no data on the separate tests. Altogether, we have accessed GA data on more than 960,000 young men over a period of 45 years. Mostly due to a greater number of seamen, the proportions having intelligence test scores were somewhat lower in the older cohorts. The average proportion over all cohorts was 0.85 (S.D.=.07).

In this file, standing height data (measured as part of the medical examination) was also available. Height data for other relevant cohorts have been compiled from the Statistical Yearbook published by Statistics Norway.

Data on the separate tests have been retrieved for the draft cohorts 1957 through 1959 and 1993 through 2002 (a small number from 2003 was included in the 2002 cohort). Altogether, subtest data was available for approximately 210,000 conscripts. In the 1957–1959 draft cohorts test data were available for 80–85% (about 52,000) of the men eligible for drafting these years. Data for the draft cohorts 1993–2002 comprised about 55% ($N \approx 158.000$) of the young men appearing before the draft board during this period. In Norway, the drafting is done separately in several regions, and data from all the regions for all the cohorts was not possible to retrieve. In the 1993 and 1994 cohorts, there were data sets from three and five out of seven regions, respectively. Data from 1995–1997, inclusive, stemmed from four regions. Data sets from the whole country were available for the cohorts from 1998 through 2000, and for the two most recent draft cohorts, data sets from six and four regions, respectively, were retrieved. On the average, about 85% of the cohorts from the relevant regions had intelligence test data.

2.3. Scales and norms

The data sets from 1957 and 1993 through 2002 contained raw scores on the tests. In the 1957–1959 data, GA in stanine scores (M=5, S.D.=2) according to the 1954 norms on all three tests were available. The GA scores in the 1969–2001 draft cohorts were also given on a stanine scale. In the present paper, scores have often been transformed into IQ equivalents (M=100, S.D.=15).

In the analyses of data, both raw score statistics and statistics on scaled scores have been used. When scaled scores have been analyzed, the test norms from 1954 have been used as the reference point for comparisons. In all the analyses applying raw scores (either in the analyses of the raw scores themselves or as basis for scaling), zero scorers have been excluded. Raw scores on the Word Similarities and Figures Tests were scaled according to the 1954 norms directly. In connection with the modernization of the contents of the Arithmetic test in 1963, stricter norms on this test were introduced. We had access only to the 1963 norms for this particular test. Flynn (1987, p. 174–175), who included Norwegian data (cf. Rist, 1982) in his review of Western intelligence test scores, estimated that this norm change corresponded to 7.5 IQ points. This is probably a slight overcorrection. In the 1957 data set, we calculated GA in stanine units from the raw scores according to the 1963 norms for the Arithmetic test and the 1954 norms on the other two tests (Figures and Word Similarities). The GA scores calculated in this manner were compared with the GA scores according to the 1954 data on all three tests. The difference between the mean GA scores was 0.28 stanine points (2.1 IQ points), corresponding to 6.3 IQ points on the Arithmetic test. The scaled scores on the Arithmetic test have accordingly been elevated by this amount in the data sets from 1957 and 1993–2002. In the data set comprising the 1969–2001 draft cohorts, GA scores were corrected upwards by 2.1 IQ points.

In the 1958 and 1959 draft cohorts, the scores on each test were given on an 11-point scale according to norms from 1961, which have been unavailable to us. The 1954 score equivalents were calculated by taking the differences between the 1954- and 1961-normed scores in the 1957 data set. These differences were added to the means in the 1958 and 1959 data sets.

The test battery was restandardized in 1980 on the basis of data from 1974 (Storsve, 1975). The mean GA scores according to the new norms were lowered 1.02 stanine points relative to the 1954/1963 norms. In the 1969–2001 draft cohorts, where only GA scores were available, we have corrected these upwards by 7.5 IQ points for the draft cohorts from 1980 and later.

3. Results

Fig. 1 displays the secular trends in GA mean scores (IQ units) from 1954 (IQ mean set to 100) through 2002.

The GA means seem to have increased more or less linearly from the 1954 to the 1969 draft cohort. In this period, the gain in mean GA was 8.6 IQ points, corresponding to an average gain of nearly 0.6 IQ points each year. From 1970 to 1976, inclusive, the gain was 1.4 IQ points (approximately 0.2/year). From 1978 to the beginning of the 1980s, there was a remarkable decline in the mean GA scores, corresponding roughly to 1.2 IQ points, which is almost the whole increase during the period 1970–1976. From the beginning of the 1980s to the mid 1990s there was a more or less steady increase in the GA means amounting to approximately 3 IQ points, corresponding to a gain of about 0.2 IQ points each year. From the mid 1990s or so, the GA means were declining again. During the whole period 1954–2002, the gain has been 10.8 IQ points, or an average yearly increase of 0.23 IQ points, which is on the low side compared with the 0.3 gain rate found in many other industrialized countries.

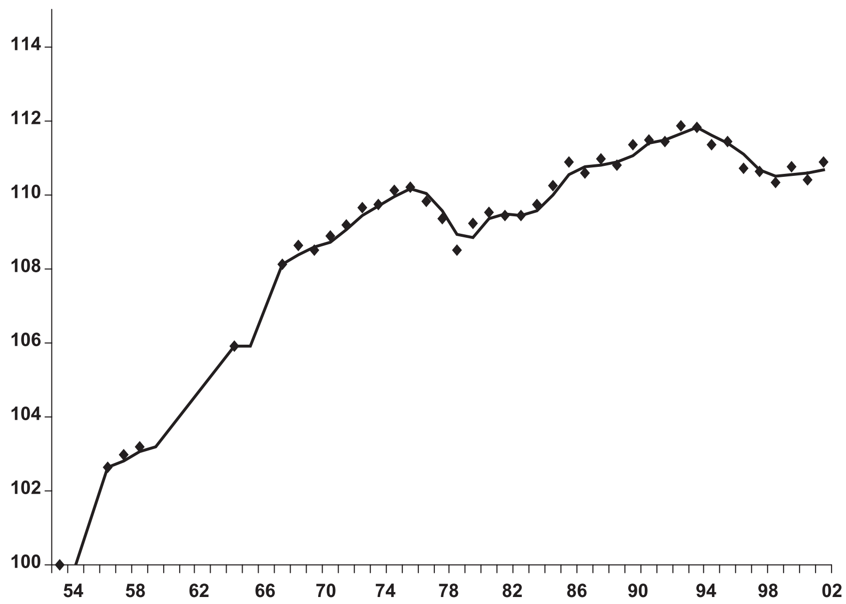

Fig. 1. GA of Norwegian conscripts (in IQ score units) by year of testing [data for 1965 and 1968 (corrected) have been adopted from Flynn, 1987, Table 4].

The time trends in raw scores (means, standard deviations, skewness, and kurtosis) on the Arithmetic, Figures, and Word Similarities for the draft cohorts 1957, 1968, 1974, 1977, and 1993 through 2002 are shown in Table 1. The data for 1968 and 1977 are from Rist (1982; cf. Flynn, 1987) and the 1974 data are from Storsve (1975).

The mean raw scores on the Arithmetic test increased steadily from 1957 to about the end of the 1960s. Keeping in mind that the content of Arithmetic test was slightly altered in 1963, the change in mean scores between the 1957 and 1968 draft cohort (1.3 raw score points) may be interpreted as an increase in arithmetic ability in this period. A peculiar and substantial drop of the means occurred from about 1968, followed by slight gains from the late 1970s to the early 1990s. A more or less steady decline of the Arithmetic mean scores was apparent in the draft cohorts 1993–2002. Notably, the highest mean was in the 1968 draft cohort. The gains in Word Similarities mean scores seemed to last until the mid 1970s. In the period from the mid 1970s to the early 1990s, the means remained more or less the same, followed by quite steady decline until 2002. The mean scores on the Figures Test behaved differently. There was a quite steady increase in mean scores on this test until the late 1990s. No systematic changes were apparent from the late 1990s to 2002. Possible ceiling effects (see below) may have depressed the means of the Figures Test to some extent.

To get a clearer picture of the relative changes of the three tests, the raw scores have been transformed to F scores and then to IQ scores (Fig. 2). The 1954 norms have been used as reference in Fig. 2, allowing the inclusion of data from the 1958 and 1959 cohorts. Only summary statistics on the raw scores were available for the test years 1968, 1974, and 1977. For these years, data from Flynn (1987, Table 4) with corrected Arithmetic scores was adopted. An extra bonus was the inclusion of data from 1963 (on Arithmetic) and from 1980, not available otherwise.

Table 1
Mean, standard deviation, skewness, and kurtosis of the raw scores on the Arithmetic, Figures, and Word Similarities tests by year of testing

Test year	Arithmetics					Figures					Word Similarities				
	M	S.D.	Skew	Kurt	*n*	M	S.D.	Skew	Kurt	*n*	M	S.D.	Skew	Kurt	*n*
1957	17.50	5.46	−.20	−.64	16792	23.11	5.44	−.80	.64	16247	27.26	11.95	.06	−.83	16453
1968[a]	18.83	5.41	–	–	788	25.88	4.28	–	–	788	31.70	11.08	–	–	788
1974[b]	18.30	5.20	–	–	5502	26.20	4.30	–	–	5513	33.50	10.20	–	–	5510
1977[a]	17.95	5.31	–	–	718	26.20	4.46	–	–	718	33.36	10.00	–	–	718
1993	18.54	4.64	−.28	−.34	14752	27.94	3.89	−1.57	5.37	14755	33.48	9.08	−.37	−.32	14750
1994	18.17	4.70	−.28	−.31	16990	27.87	3.99	−1.65	5.80	17006	33.10	9.00	−.32	−.42	17000
1995	18.00	4.65	−.16	−.47	11571	27.91	3.85	−1.42	4.67	11576	32.66	9.00	−.27	−.39	11546
1996	18.10	4.59	−.16	−.47	12807	28.03	3.75	−1.40	4.85	12801	32.82	8.87	−.27	−.39	12795
1997	17.73	4.66	−.16	−.44	13110	28.04	3.78	−1.50	5.38	13114	32.30	8.93	−.29	−.34	13113
1998	17.72	4.65	−.15	−.45	25342	27.92	3.96	−1.61	5.76	25346	32.31	9.00	−.32	−.24	25333
1999	17.82	4.58	−.18	−.35	21941	27.77	4.10	−1.64	5.60	21953	32.14	8.94	−.34	−.25	21947
2000	17.95	4.60	−.21	−.39	16644	27.96	3.94	−1.56	5.37	16640	32.45	8.81	−.32	−.29	16640
2001	17.69	4.67	−.18	−.45	11733	27.91	4.01	−1.60	5.55	11738	32.26	8.95	−.33	−.28	11726
2002	17.45	4.67	−.17	−.40	13405	27.98	4.06	−1.76	6.39	13414	32.12	9.03	−.32	−.31	13409

[a] From Rist (1982).
[b] From Storsve (1975).

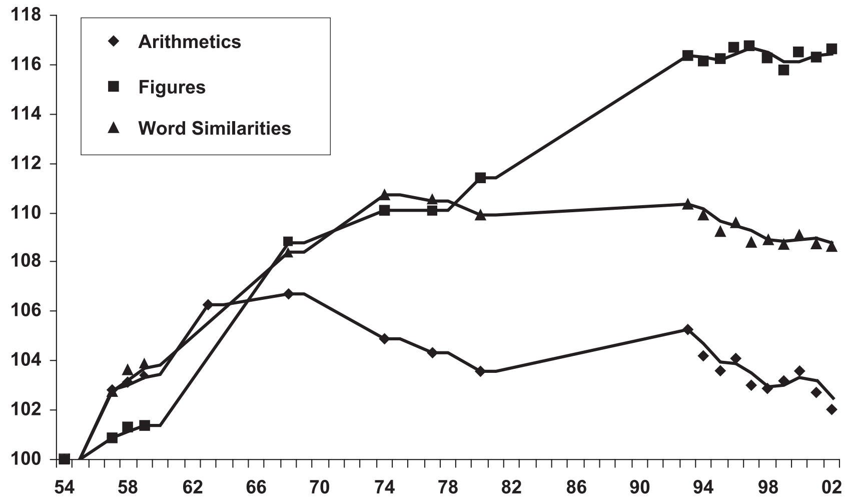

Fig. 2. Scores on each of the three ability tests of Norwegian conscripts (in IQ units) by year of testing [the data points for 1963, 1968, 1974, 1977, and 1980 (with corrections of the Arithmetic scores) have been adopted from Flynn, 1987, Table 4].

Fig. 2 confirms the impression that all the test means increased more or less in tandem until the late 1960s–early 1970s (test years). Thus, the relatively large gains on the GA means in this period (approximately 0.6 per year, cf. Fig. 1) were due to gains on all the tests. It seems also clear that the decrease in the GA means from the mid 1970s to the early 1980s is mainly caused by the decreasing means on the Arithmetic test, together with slightly declining mean scores on the Word Similarities Test from 1974 to 1980. The increase of GA means in the draft cohorts from the early 1980s to the mid 1990s was almost exclusively due to gains on the Figures Test (with due consideration of missing information in the years between). The decline of the GA means from the mid 1990s was apparently mainly caused by a relatively sharp decline in both the Arithmetic and Word Similarities means. During the whole period from 1954 to 2002, the gains on the Arithmetic, Word Similarities, and Figures Tests were 2.5 (0.05 IQ points per year), 9.6 (0.2 IQ points per year), and 17 (0.35 IQ points per year) IQ points, respectively. The general picture seems to be a quite steady but decreasing gain rate of the means on the Figures Test until the mid to late 1990s, but no gains later. The mean gains on the Arithmetic test scores came to a more or less complete stop somewhere in the 1960s, followed by a period of decline until about 1980. Small gains until 1993 were followed by a period of more or less steady decline. The Word Similarities Test showed steady mean gains until the mid 1970s, followed by a small decline until about 1980. The means on this test were quite unchanged from about 1980 to 1993, and declined for the rest of the observation period.

Mean changes were accompanied by changes in standard deviations, skewness, and kurtosis (Table 1). The scores on all three tests showed decreasing trends in the standard deviations in the 1993–2002 draft cohorts relative to the 1957 cohort (Table 1). The standard deviation in the 1993–2002 cohorts was about 85% of the standard deviation on the Arithmetic test. The corresponding numbers for Figures and the Word Similarities were 73% and 75%, respectively. The score distributions on the Figures and the Word Similarities tests both showed increasing skewness from the 1957 to the 1993–2002 conscript cohorts. These changes may either be due to ceiling effects or real-world changes of the score distributions, or both. All the means were larger than 50% of the maximum obtainable scores (15, 18, and 27 for the Arithmetic, Figures, and Word Similarities, respectively). The mean of the Arithmetic test did not increase much (from 17.5 in the 1957 cohort to 17.9 in the 1993–2002 cohorts). The means of the 1968 and 1974 cohorts were higher than the 1993–2002 mean, and so were the standard deviations (Table 1). Similarly, the means of the Word Similarities scores in the 1974 and 1977 cohorts were of the same order of magnitude as in the 1993–2002 cohorts, but the standard deviations were larger. This pattern does not indicate appreciable ceiling effects in the 1993–2002 cohorts relative to the 1957 cohort with regard to the Arithmetic and Word Similarities tests. The mean score on the Figures Test in the 1957 cohort was 66% of the maximum score, increasing to 72–73% in the 1968–1977 cohorts and 77.5% in the 1993–2002 cohorts, and increasing means seem to be associated with decreasing standard deviations (Table 1). Seen in isolation, this pattern may indicate ceiling effects in the more recent cohorts relative to the older ones. On the other hand, the relative decreases of the standard deviations of the Figures scores were of the same order of magnitude as for the Word Similarities scores, indicating that, at least, some of the standard deviation changes of the Figures standard deviations are due to real-world distribution changes.

We have addressed the question concerning whether the mean changes have been unevenly distributed over different ability levels by calculating the mean scores below and above the median of the 1957–1959 (only 1957 for raw scores) and the 1993–2002 groups and studied the difference between corresponding means in the two cohort groups. The results of this analysis are shown in Table 2.

It is clear from Table 2 that the gains over cohort groups tended to be largest below the median. For instance, with regard to the Word Similarities Test, the difference between the mean scores of the 1993–2002 and the 1957–1959 draft cohorts below the median was about 11 IQ points (8.5 raw score points). Above the median, the corresponding difference was about 3 IQ points (2.4 raw score points). The corresponding differences also decreased on the Figures Test, but not to the same extent (about 19 IQ points below the median and 12 IQ points above the median). Due to possible ceiling effects on this particular test, the gains in means above the median may, to some extent, have been suppressed. With regard to the Arithmetic test, there was a gain in the mean scores below the median, and a slight loss in the above median mean scores in the 1993–2002 cohorts relative to the 1957–1959 cohorts, indicating a lower prevalence of both low and high scorers in the more recent cohorts.

Table 2
Mean scores in IQ units and raw scores (in brackets) below and above the median for the Arithmetic, Figures, and Word Similarities tests in the 1957–1959 (1957) and 1993–2002 draft cohorts

Test year	Mean scores below median			Mean scores above median		
	Arith	Figures	Word sim	Arith	Figures	Word Sim
1957–1959	90.7 (13.5)	87.4 (19.3)	90.6 (17.6)	115.8 (22.5)	114.9 (27.6)	115.9 (37.6)
1993–2002	93.6 (14.3)	106.5 (25.1)	101.5 (26.1)	114.5 (21.9)	126.3 (30.8)	119.0 (40.7)
Difference	2.9 (0.8)	19.1 (5.8)	10.9 (8.5)	−1.3 (−0.6)	11.4 (3.2)	3.1 (3.1)

Arith=arithmetic.
Word Sim=word similarities.

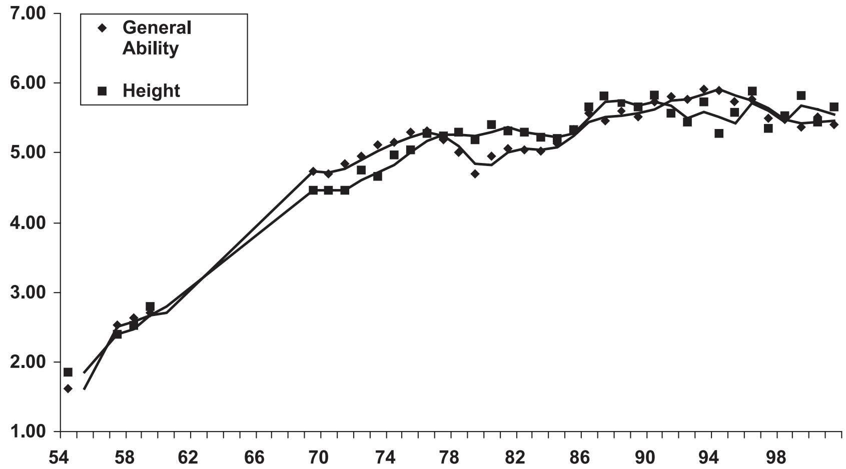

Fig. 3. Mean standing height and mean GA (both in z scores units+5) by year of testing.

The mean standing height among Norwegian conscripts increased from about 176.5 cm from the mid 1950s (Statistical Yearbook, 1960) to about 178.8 cm in 1969. The height means further increased to about 179.6 cm in 1987. From about 1987 or so, no systematic change was apparent. Standard deviations of the height distributions increased slightly in the draft cohorts 1969–2002. Fig. 3 shows the means of standing height and GA (both transformed to z scores in an aggregated file, and added 5 to remove negative numbers) according to draft year.

It can be seen from Fig. 3 that the mean standing height and mean GA follow each other quite closely. The within-person correlation between standing height and GA showed a continuous decrease from .17 to .14 in the period 1969–2002.

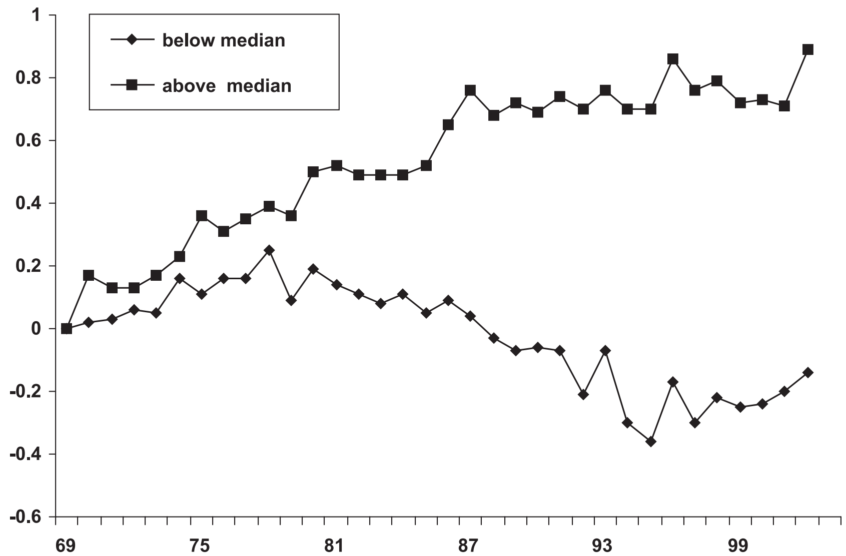

Fig. 4. Mean height below and above the median height by year of testing.

Table 3
Intercorrelations between the test scores in IQ units and raw scores (in brackets) in the 1957– 1959 (1957) draft cohorts (lower triangle) and in the 1993–2002 cohorts (upper triangle)

	(1) Arith	(2) Fig	(3) Word S
(1)	–	.54 (.56)	.57 (.60)
(2)	.64 (.68)	–	.48 (.49)
(3)	.73 (.75)	.61 (.64)	–

Arith=arithmetics.
Fig=figures.
Word S=word similarities.

Fig. 4 displays the change trends above and below the median height relative to the 1969 draft cohort. The gains above the median were at least as large as the gains below the median in all the cohorts from 1969 to 2002.

The intercorrelations between the tests in raw scores and IQ units are displayed in Table 3. It is seen that the correlations were somewhat lower in the 1993–2003 than in the 1957–1959 (1957) draft cohorts. The intercorrelations (raw scores) reported by Storsve (1975) were about midway between the correlations shown in Table 3.

The possible range restrictions of the Figures scores complicate the interpretation of the declining intercorrelations. Assuming that range restrictions are the main cause of the declines, the largest decline should be observed in the correlations between Figures and the other two tests, whereas the Arithmetic–Word Similarities correlation should remain approximately the same. This is not in accordance with the observed declines seen in Table 3. Thus, the correlation between the Arithmetic and Word Similarities in the 1993–2002 cohort declined by 22% (20% in the raw scores) relative to the 1957–1959 (1957) cohorts, whereas the correlation decline was 16% (18% in the raw scores) between Arithmetic and Figures. The observed decline in the correlation between Figures and Word Similarities scores was 21% on the IQ-unit scale and 24% on the raw score scale. This pattern indicates that most of the observed declines in the intercorrelations are real-world declines, entailing a less pronounced *g* factor in the more recent cohorts.

4. Discussion

The tests used to assess the intellectual ability among Norwegian conscripts are representative of subtests regularly included in standard intelligence tests. Thus, the mathematics and language tests are similar to the subtests in WAIS, and the nonverbal test was explicitly constructed to be similar with the Raven Progressive Matrices Test. In the Cattell (1987) system, the first two tests measure crystallized intelligence, whereas the last one measures fluid intelligence. The GA score, obtained by combining the scores on the three tests, correlates substantially (in the .70s) with the WAIS IQ scores. The GA scores reported in the present paper are thus quite comparable with the scores obtained on standard intelligence tests.

Test scores have been attained from very large samples comprising a large proportion of Norwegian (male) conscripts. The selection due to the exclusion of physically or psychologically disabled has probably been more or less the same from year to year and is not likely to affect the observed secular trends in intelligence test scores.

The potentially most serious threat to the present results seems to be the possible ceiling effects on the Raven-like Figures Test that may have suppressed the means and standard deviations of the scores in newer cohorts relative to older ones.

Earlier reports have monitored secular trends up to the birth cohorts around 1980 (Colom, Juan-Espinosa, & Garciá, 2001; Emanuelsson et al., 1993; Teasdale & Owen, 2000). The present results extend the time window to the birth cohorts in the mid 1980s. The mean scores of all the three tests show quite substantial gain rates in the conscript cohorts from the mid 1950s to the late 1960s–mid 1970s (Fig. 2 in the present paper, and Table 4, Flynn, 1987), and the mean GA scores increased accordingly in this period (Fig. 1).

The peculiar drop in Arithmetic mean scores in the conscript cohorts from about the end 1960s to about mid 1970s (Fig. 2 and Table 1) was so extensive that it caused a drop in the GA scores (Fig. 1). Rist (1982) convincingly argued that this particular decrease is due to teaching of modern mathematics (more algebra) at the expense of training in arithmetic operations. This program was terminated after a few years. Despite some gains from about 1975 to 1993, the means on the Arithmetic test never quite reached the 1968 level. The means of the Word Similarities Test peaked in the mid 1970s and did not change much from then to 1993. From about 1993 to 2002, both the Arithmetic and the Word Similarities Test scores showed declining means.

The means on the Raven-like Figures Test (presumably measuring fluid intelligence) increased from the mid 1950s to the late 1990s, although with decreasing gain rates from the mid 1970s. The total gain during this period was about 16–17 IQ points. In accordance with earlier results (cf. Flynn, 1999), we found smaller gains on the two tests measuring crystallized intelligence (Fig. 2). The mean gains on the Raven-type test seem to have ceased somewhere in the late 1990s (Table 1 and Fig. 2). As far as we know, this is the first observation indicating a cessation of secular gains on a Raven-type test. Considering the recent observations of decreasing gains in Denmark (Teasdale & Owen, 2000), it seems probable that the end of the period of secular gains may be imminent also in that country. In Sweden, a leveling and even possible slight declines of the means on visuospatial, reasoning, and verbal tests were observed in the period from 1985 to 1990 among 13-years-olds (Emanuelsson et al., 1993). The secular trends may seem to behave quite similarly in the Scandinavian countries. Considering the similarities of the political, educational, and social systems of these countries, this may not be entirely coincidental.

The present results indicate that the secular trends in Norway may have been unevenly distributed over different ability levels. This effect is especially clear on the verbal test, where the mean increase below the median score from the 1957–1959 to the 1993–2002 draft cohorts was about three times the increase above the median score (Table 2). On the Raven-like test, the gains below the median outsized the gains above by a factor of about 1.5. Possible ceiling effects may have suppressed increases above the median on this particular test. The results on the mathematics test indicate lower prevalence of both low and high scorers. The increase below the median was somewhat larger than the decrease above and allowed a slight net mean gain in the 1993–2002 cohorts relative to the 1957–1959 cohorts.

Unfortunately, we do not have access to the full distribution of scores for cohorts between the 1957–1959 and the 1993–2002 cohorts. However, data on conscripts in five birth cohorts from 1939–1969 (tested in the period 1957–1987) in Denmark show that almost all the mean changes are due to the disappearance of low scorers (Teasdale & Owen, 1989). Decreasing variances, which may be an indication of unevenly distributed changes, have been observed in several countries, for example, the Netherlands (Flynn, 1987) and Israel (Flynn, 1998a). It is certainly possible that all or parts of these decreases are due to ceiling effects, but not necessarily so. We find ourselves in agreement with Teasdale and Owen (1989), who point out that assuming unchanging variances that seems to have been done by some researchers (e.g., Flynn, 1987), may be a dubious practice. In a similar vein, routinely normalizing the scores and setting the standard deviation in every restandardization of intelligence tests to a fixed value (most often 15 IQ points) may mask real-world changes of the intelligence distribution in a population.

The decreasing intercorrelations between the tests towards the more recent cohorts found in the present results (Table 3) and in other studies (Kane & Oakland, 2000; Lynn & Cooper, 1993) may indicate a change in the factor structure of the test scores over the 40–50 years studied in the present paper. It seems that the celebrated *g* factor (Jensen, 1998; Sternberg & Grigorenko, 2002) has less impact than it used to have. In itself, this is no wonder. After all, the *g* factor has been derived from the pattern of covariations between abilities in a certain society in a given time period.

At first glance, the quite close association between increases in mean standing height and mean intelligence test scores found in the present study from the draft cohorts in the mid 1950s to about 1987 (Fig. 3) may seem to suggest that nutrition improvements (and better health care) may have been a potent factor in this period. However, most of the height gains seem to be due to increases in the upper half of the height distribution (Fig. 4). This is quite the opposite of the tendencies of intelligence test scores. These results seem to weaken the nutrition theory considerably, at least for the birth cohorts after 1950 (test year 1969). A further blow to the nutrition theory is the fact that intelligence score gains continued for about a decade after the cessation of height gains. It is still possible that nutrition and health care may have been among the effective factors in the birth cohorts before the 1950s and in less developed nations (cf. Daley et al., 2003).

Further insight into the causes of the Flynn effect may be attained by studying putative causes in countries where the secular gains are in different phases. Promising candidates may be factors connected to improved conditions of life, like family size, leisure time, changing rearing styles (Baumrind, 1991; Bronfenbrenner & Ceci, 1994), and use of day-care facilities (Hartmann, 1991). The impact of computer games and growing access to mass media may also be of some importance. Studies aimed at understanding the types of cognitive competencies involved in different types of tests (Carpenter, Just, & Shell, 1990) and neuropsychological studies (Gray, Chabris, & Braver, 2003) seem to have great explanatory potential. In addition, experimental and other types of studies investigating potentially effective factors in changing those abilities could be enlightening.

Acknowledgements

The authors are indebted to the Norwegian Armed Forces and the Norwegian Social Science Data Services for kindly making the data used in the present paper available to us. Special thanks to Olav Storsve at the Psychological Services, Norwegian Armed Forces, for assistance in our efforts to ascertain data.

References

Baumrind, D. (1991). Parenting style and adolescent development. In J. Brooks-Gunn, R. Lerner, & A. C. Petersen (Eds.), *The encyclopedia of adolescence* (pp. 746–758). New York: Garland Press.

Bronfenbrenner, U., & Ceci, S. J. (1994). Nature–nurture reconceptualized in developmental perspective: A bioecological model. *Psychological Review*, *101*, 568–586.

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices. *Psychological Review*, *97*, 404–431.

Cattell, R. B. (1987). *Intelligence: Its structure, growth and action*. Amsterdam: North-Holland.

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

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.

Detterman, D., & Daniels, M. H. (1989). Correlations of mental tests with each other and with cognitive variables are highest for low IQ groups. *Intelligence*, *13*, 349–359.

Emanuelsson, I., Reuterberg, S. E., & Svensson, A. (1993). Changing differences in intelligence? Comparisons between groups of 13-year-olds tested from 1960 to 1990. *Scandinavian Journal of Educational Research*, *37*, 259–276.

Emanuelsson, I., & Svensson, A. (1990). Changes in intelligence over a quarter of a century. *Scandinavian Journal of Educational Research*, *34*, 171–187.

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

Flynn, J. R. (1998a). Israeli military IQ tests: Gender differences small, IQ gains large. *Journal of Biosocial Science*, *30*, 541–553.

Flynn, J. R. (1998b). 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. (1999). Searching for justice. The discovery of IQ gains over time. *American Psychologist*, *54*, 5–20.

Gray, J. R., Chabris, C. F., & Braver, T. S. (2003). Neural mechanisms of general fluid intelligence. *Neuroscience*, *6*, 316–322.

Hartmann, E. (1991). Effects of day care and maternal teaching on child educability. *Scandinavian Journal of Psychology*, *32*, 325–335.

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

Kane, H., & Oakland, T. H. (2000). Secular declines in Spearman's *g*: Some evidence from the United States. *Journal of Genetic Psychology*, *16*, 337–345.

Lynn, R. (1990). The role of nutrition in secular increases of intelligence. *Personality and Individual Differences*, *11*, 273–286.

Lynn, R., & Cooper, C. (1993). A secular decline in Spearman's *g* in France. *Learning and Individual Differences*, *5*, 43–48.

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

Martorell, R. (1998). Nutrition and the worldwide rise in IQ. In U. Neisser (Ed.), *The rising curve. Long term gains in IQ and related measures* (pp. 183–206). Washington, DC: American Psychological Association.

Neisser, U. (Ed.). (1998). *The rising curve. Long term gains in IQ and related measures* (pp. 183–206). Washington, DC: American Psychological Association.

Rist, T., 1982. *The intellectual development in the population in the light of the social development* (In Norwegian). Thesis, Oslo University.

Statistical yearbook, 1960. Oslo: Statistics Norway.

Sternberg, R. J., & Grigorenko, E. L. (Eds.). (2002). *The general factor of intelligence. How general is it?* London: Lawrence Erlbaum Associates.

Storsve, O. (1975). *Report on the re-standardization of the Norwegian Army intelligence tests*. London: Psychological Services of the Norwegian Army In Norwegian.

Sundet, J. M., Tambs, K., Magnus, P., & Berg, K. (1988). On the question of secular trends in the heritability of IQ test scores: A study of Norwegian twins. *Intelligence*, *12*, 47–59.

Teasdale, T. W., & Owen, D. R. (1989). Continuing secular trends 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.