Intelligence 37 (2009) 25–33

Contents lists available at ScienceDirect

Intelligence

Comparability of IQ scores over time ☆

Olev Must [a,*], Jan te Nijenhuis [b], Aasa Must [c], Annelies E.M. van Vianen [d]

[a] University of Tartu, Department of Psychology, Tiigi 78, Tartu 50410, Estonia
[b] University of Amsterdam, Work and Organizational Psychology, the Netherlands
[c] Estonian National Defence College, Estonia
[d] University of Amsterdam, Work and Organizational Psychology, the Netherlands

ARTICLE INFO

Article history:
Received 4 July 2007
Received in revised form 15 February 2008
Accepted 9 May 2008
Available online 20 June 2008

Keywords:
IQ test
Comparability
Secular score gains
Flynn Effect
Estonia
National Intelligence Test

ABSTRACT

This study investigates the comparability of IQ scores. Three cohorts (1933/36, 1997/98, 2006) of Estonian students (N=2173) are compared using the Estonian National Intelligence Test. After 72 years the secular rise of the IQ test scores is.79 SD. The mean .16 SD increase in the last 8 years suggests a rapid increase of the Flynn Effect (FE) in Estonia. The measurement is not strictly invariant, which means that the IQ scores of different cohorts are not directly comparable. Less than perfect comparability of test scores is caused by at least two factors: time between measurements and societal/educational changes between cohorts. As was to be expected, the meaning of subtests and the meaning of the *g* score have changed over time.

1. Introduction

The populations of several countries have increased in average IQ by about 3 points a decade over the last 60 years (Flynn, 2007), and this development is known as the Flynn Effect (FE). Recent papers from Scandinavian countries prove that the effect has come to a standstill and it even appears that a negative Flynn Effect (Lynn & Harvey, 2007) is emerging. The present study focuses on the question of whether the FE is continuing in Estonia.

There is no consensus about the causes of rising IQ scores, but there is evidence that IQ scores change their meaning over time: IQ scores from different time periods are not perfectly comparable (see Jensen, 1998; Neisser, 1998; Colom, Juan-Espinosa, & García, 2001; Wicherts et al., 2004; Flynn, 2007). The second question in this study is how strongly IQ scores change their meaning over time. We analyze old and recent data on the Estonian National Intelligence Test (NIT) with 72 years in between; this is the largest time interval in the FE literature.

1.1. Are secular gains continuing?

Recent studies show IQ scores rising in less-developed parts of the world, for example in Kenya (Daley, Whaley, Sigman, Espinosa & Neumann, 2003) and in the Caribbean (Meisenberg, Lawless, Lambert & Newton, 2006). There is also recent evidence of IQ test scores continuing to rise in industrialized countries (e.g. in Argentina and the US, see Flynn, 2007). The first study suggesting the end of the increase in IQ scores was by Emanuelsson and Svensson (1986) (see also Emanuelsson, Reuterberg, and Svensson, 1993). Recent papers offered additional data demonstrating that the test scores in Scandinavian countries are no longer increasing and even suggesting a decline of IQ scores (Sundet, Barlaug & Torjussen, 2004; Teasdale & Owen, 2005; Shayer, Ginsburg & Coe, 2007; Teasdale & Owen, 2007).

1.2. Are the gains on g?

Several causes of the secular gains in test scores have been suggested: nutrition (Lynn, 1990, 1998); improvements in health care, increased outbreeding and heterosis (Jensen, 1998; Mingroni, 2004, 2007); education and changes in social environment (Teasdale & Owen, 1987; Ceci, 1991; Jensen, 1998;

☆ The preparation of this paper was supported by the Estonian Scientific Foundation (Grant 5856).
* Corresponding author. Tel.: +372 7 375 912; fax: +372 7 352 900.
E-mail address: Olev.Must@ut.ee (O. Must).

0160-2896/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.intell.2008.05.002

Dickens and Flynn, 2001, 2002), and changes in fertility patterns (Sundet, Borren & Tambs, 2007). Jensen (1998, p. 332) states that the most reasonable hypothesis to account for the secular trend in IQ is that the IQ increments consist of two main parts: (1) a functional, *g*-loaded part due to the secular trend in biological environmental improvements that produce general biological effects, and (2) a part that is largely "hollow" with respect to *g* and is slightly, if all, reflected in a functional increase of real-life problem solving ability due to a secular trend in non-biological environmental effects. Jensen (1998) suggests that less than half of the FE constitutes a real gain on *g*, while more than half of the effect is on broad abilities, narrow abilities, and test-specific abilities and so of limited generalizability or even non-generalizable (i.e., "empty").

Flynn's writings (Dickens & Flynn, 2001, 2002; Flynn, 1999, 2000, 2007) seem to suggest that there has been no gain on the *g* factor: brain power remains essentially unchanged over generations. All the gains are on broad, narrow, and test-specific abilities, due to changing educational and work requirements. Flynn emphasizes that the gains on individual IQ tests are real, because they translate into improved real-world behavior.

1.3. Measurement invariance

The US gains in IQ scores are about two standard deviations (Flynn, 2007). Strict measurement invariance would mean that there would be no problem using norm tables from the 1950s in 2008, which would imply that half of the present US youth would be able to perform a job of high complexity. This is clearly not the case, so a mean IQ score of 130 does not have the same meaning in 2008 as in the 1950s, and lack of measurement invariance is what one would expect. Indeed, lack of measurement invariance is implicit in all the papers on the Flynn Effect, and Wicherts et al. (2004) showed that in every dataset on the FE they studied, including the NIT, there was lack of measurement invariance. This means that IQ scores have different meanings in different cohorts. It is important to pinpoint the sources of violation of the measurement invariance and find their causes.

Spearman's principle of the indifference of the indicator means that *g* or general mental ability can be measured using a large variety of measures, and implies that *g*-s extracted from different test batteries should be highly correlated (Johnson, Bouchard, Krueger, McGue, & Gottesman, 2004; Johnson, te Nijenhuis & Bouchard, 2008). For the NIT, we therefore also expect strong general factors to be present in all cohorts. However, these general factors most likely also will show lack of measurement invariance, meaning that *g* scores of various cohorts cannot be directly compared.

1.4. Research questions

In the current paper we focus on two research questions.

1. A previous study (Must et al., 2003) showed a clear FE in comparing cohorts from 1933/36 and 1997/98; our next question is whether the scores are still increasing. We compare cohorts from 1933/36 and 2006, and 1997/98 and 2006, respectively.
2. Previous research on (Estonian) NIT data showed lack of measurement invariance (Wicherts et al., 2004). Using a more solid database, now based on data from three cohorts, we try to find the causes of violation of invariance.

2. Method

2.1. Measures

The American National Intelligence Test (NIT) (Terman, 1921) was adapted into Estonian at the beginning of the 1930s by Tork (1940). The Estonian version of the NIT consists of an A and a B scale and there are five item types (a description of the scales see Must et al., 2003). Scales A and B both use the same five item types, so every item type is used twice. In this paper we use the following abbreviations:

A1: Arithmetic (16 items requiring a solution for one unknown quantity).
A2: Sentence Completion (20 items requiring filling in missing words to make sentence understandable and correct).
A3: Concept Comprehension (24 items requiring selecting two characteristic features from among those given. For example "Cat: sphere, claws, eating, eyes, mouse?").
A4: Synonyms–Antonyms (40 items requiring evaluation of whether the words presented mean the same or opposite).
A5: Symbol–Number Correspondence (120 items requiring decision which digit must be assigned to a present symbol according to a key).
B1: Computation (22 items requiring addition, subtraction, multiplication, and division of both integers and fractions).
B2: Information (40 items requiring picking the correct answer from an array of questions. For example "What is the first month of the year?").
B3: Vocabulary (40 items requiring knowledge about the qualities of different objects, "Are books used in school?").
B4: Analogies (32 items requiring transferring the relation between two given words to other presented words).
B5: Comparisons (50 items requiring judgment about sameness of sets of numbers, family names, and graphic symbols presented in two columns).

In the current replication study the original scales from the 1930s are used with only minimal changes, for instance, in the currency units; moreover, some of the facts in the Information subtest were brought up to date, such as World War I being changed to World War II. Scales come with exercise parts on separate sheets of the same test booklet. The first practice items are given to the whole group and then test-takers practice independently. All scales are speeded. As a rule, the time for practice was 30 s per subtest, and for taking the actual test the time varied from 2 min (Comparisons) to 4 min (Computation). As the NIT is a group test and there are 20 different timing episodes, it requires attention and motivation from the test-takers. The 1930s NIT (Tork, 1940) required linear transformation of the subscales to calculate the composite score.

The NIT is an old test, based on the IQ theories and measurement traditions that dominated in the 1920s. The NIT has been replaced by other tests that assess mental abilities. However, for the purpose of the present study it is the only source to compare IQ data from different periods.

Table 1
Secular gains in subtests (in SDs) during the period 1933/36–2006 (12–13-year-old students, respectively from 4th grade in 1933/36 and from 6th grade in 2006)

Subtest	1933/36 (*N*=270)		2006 (*N*=243)		Gains
	Raw mean	SD	Raw mean	SD	
A1: Arithmetic	13.0	4.3	15.2	4.4	.51
A2: Sentence Completion	19.6	5.1	29.6	5.9	1.81
A3: Concepts	22.9	6.7	34.8	7.4	1.68
A4: Synonyms–Antonyms	19.9	9.1	31.3	6.3	1.44
A5: Symbol–Number	18.5	4.9	30.5	5.3	2.33
B1: Computation	20.9	4.1	21.5	4.1	.15
B2: Information	16.3	5.4	21.0	4.5	.94
B3: Vocabulary	21.7	5.1	25.3	5.9	.65
B4: Analogies	9.8	5.1	19.6	5.7	1.81
B5: Comparison	22.4	4.8	35.0	5.9	2.34
Mean					1.37
Increase per decade					.20

Note: sample from 1997/98 does not include 6th grade students.

2.2. Samples

For the analyses we used three samples of Estonian schoolchildren attending schools with instruction in Estonian. In 2006 approximately 82% of students from grades 6–8 attended schools with instruction in Estonian (Statistics Estonia, 2007). The presentation of results from different periods is based on a comparison of students of the same chronological age attending different grades.

1. *Sample from 1933/36* (*N*=899): students from grades 4–6 (subsamples *n*=270, 222 and 407 respectively), mean age 13.4 years, mainly (75%) from the county of Tartumaa, and from the town of Tartu. The sample is based on the original data retrieved from the Estonian Historical Archive.
2. *Sample from 1997/98* (*N*=361): students from grades 7–8 (subsamples *n*=224 and 137 respectively), mean age 13.2 years, from the same region as the first sample. The second sample is highly similar to the sample used in earlier research (Must et al., 2003).
3. *Sample from 2006* (*N*=913): students from grades 6–8 (subsamples *n*=243, 343 and 327 respectively), and mean age 13.5 years, from the same region as the two previous samples.

2.2.1. Data cleaning

Cases with extreme response patterns were excluded. The mean percent of correct answers was calculated for every respondent and it was compared with the result in each subtest. If the result in some subtest was at least 1.5 times lower than the individual mean rate, a child's test booklet was checked to find the reason for the low score. The main source of low scores was incorrect completion in subtest A4 (Synonyms–Antonyms): students did not remember the instruction (which letter to use to denote similarity of words); the second reason for deletion was the case when a student obviously did not follow the testing instruction (devoted testing time to exercising instead of taking items from the actual test). Also we deleted the data from analysis, when some subtests were not completed. This cleaning process eliminated 8% of students from the initial sample from 1933/36, 5% from the initial sample from 1997/98, and 6% from the initial sample from 2006.

Tork's adaptation of the NIT had a clear target group, namely students at the end of Elementary/Basic School — grades 4, 5, and 6 in the 1930s. However, due to differences in the age children start school, students of the same age are now in the 6th, 7th, and 8th grades, respectively. The samples are divided into age–grade subgroups to allow comparisons of same-age students although they attend different grades (for instance, we compare students from the 1930s from 4th grade with contemporary 6th-graders). The correlations among the subtests for the different samples are shown in Appendix A.

2.3. Measurement invariance

Studying measurement invariance helps to find the causes of secular score gains. To estimate measurement invariance in our FE comparisons the LISREL (Jöreskog & Sörbom, 2006) Multigroup Confirmatory Analysis (MGCFA) program was used.

3. Results

3.1. Score gains

Tables 1–3 show that the NIT scores have increased from 1933/36 to 2006 in all age groups. The youngest group shows the largest secular increase — approximately .20 SD per decade (see Table 1) — whereas the oldest group shows the smallest

Table 2
Secular gains in subtests (in SDs) during the period 1933/36–2006 (13–14-year-old students, respectively from 5th grade in 1933/36 and from 7th grade in 1997/98 and 2006)

Subtest	1933/36 (*N*=222)		1997/98 (*N*=224)		2006 (*N*=343)		Gains		
	Raw mean	SD	Raw mean	SD	Raw mean	SD	1933/36 vs. 1997/98	1933/36 vs. 2006	1997/98 vs. 2006
A1: Arithmetic	14.9	4.2	14.2	4.5	15.6	4.2	−.16	.16	.32
A2: Sentence Completion	24.2	5.7	29.4	6.1	30.2	5.7	.86	1.02	.16
A3: Concepts	27.0	6.7	38.9	7.2	35.5	6.7	1.60	1.15	−.45
A4: Synonyms–Antonyms	25.0	5.9	30.1	6.9	31.7	5.9	.59	.81	.22
A5: Symbol–Number	21.8	4.8	30.1	5.1	31.3	4.8	1.58	1.68	.10
B1: Computation	22.7	4.8	22.6	4.8	21.9	4.8	−.02	−.17	−.15
B2: Information	21.4	4.8	19.3	5.1	21.6	4.8	−.37	.03	.40
B3: Vocabulary	24.9	5.8	25.0	5.6	26.1	5.8	.01	.21	.20
B4: Analogies	13.9	5.7	17.5	5.4	20.0	5.7	.65	1.07	.42
B5: Comparisons	24.8	6.2	33.6	6.4	35.8	6.2	1.34	1.76	.42
Mean							.61	.77	.16
Increase per decade							.06	.11	. 20

Table 3
Secular gains in subtests (in SDs) during the period 1933/36–2006 (14–15-year-old students, respectively from 6th grade in 1933/36 and from 8th grade in 1997/98 and 2006)

Subtest	1933/36 (N=407)		1997/98 (N=137)		2006 (N=327)		Gains		
	Raw mean	SD	Raw mean	SD	Raw mean	SD	1933/36 vs. 1997/98	1933/36 vs. 2006	1997/98 vs. 2006
A1: Arithmetic	17.3	4.3	15.4	4.3	17.0	4.6	−.44	−.07	.37
A2: Sentence Completion	28.2	5.8	31.3	5.3	32.0	5.4	.56	.68	.12
A3: Concepts	31.7	7.2	40.3	5.9	36.5	6.3	1.31	.71	−.60
A4: Synonyms–Antonyms	30.4	7.4	30.6	7.0	33.2	5.5	.03	.43	.40
A5: Symbol–Number	25.1	6.0	30.9	4.6	32.2	4.6	1.10	1.30	.20
B1: Computation	25.3	4.9	22.9	4.7	22.0	4.7	−.50	−.69	−.19
B2: Information	26.7	6.2	20.2	4.9	23.3	5.2	−1.20	−.59	.61
B3: Vocabulary	26.5	5.2	26.2	5.9	27.6	5.4	−.05	.21	.26
B4: Analogies	15.9	5.8	18.6	5.2	20.2	6.0	.49	.73	.24
B5: Comparisons	28.5	6.5	34.9	5.0	37.1	6.2	1.1	1.4	.30
Mean							.24	.41	.17
Increase per decade							.04	.06	.20

increase — .06 SD per decade (see Table 3). The general increase of IQ scores during 72 years is .79 SD (≈1.65 IQ points per decade). The eight years between 1997/98 and 2006 yield a gain of .16 SD — approximately 3 IQ points per decade — which is twice the size of gain per decade between 1933/36 and 2006. In the period 1997/98 there are no differences in gains between the age groups studied.

At the subtest level the test-score changes are different, ranging from decreases in subtests A1 and B1 (Arithmetic and Computation, respectively; see Tables 2–3) to increases up to 2 SD in subtest B5 (Comparison) during the period 1933/36 to 2006 (see Table 1).

Table 4
g factor of the NIT in different decades

Subtest	Cohort 1933/36 (N=899)	Cohort 1997/98 (N=361)	Cohort 2006 (N=913)
	CFA factor loadings	CFA factor loadings	CFA factor loadings
A1: Arithmetic	.69	.59	.62
A2: Sentence Completion	.82	.75	.74
A3: Concepts	.79	.56	.67
A4: Synonyms–Antonyms	.74	.57	.66
A5: Symbol–Number	.52	.43	.45
B1: Computation	.51	.57	.57
B2: Information	.88	.78	.74
B3: Vocabulary	.67	.67	.62
B4: Analogies	.74	.70	.78
B5: Comparisons	.53	.42	.48
Mean factor loading on *g* factor	.69	.60	.63
Sum of squared loadings	4.9	3.8	4.1
Difference of eigenvalues		F_1=1.29; p<.01	F_2=1.20; p<.01 F_3=.93; p>.10
Mean intercorrelation of subtests	.48	.37	.39
g-loadedeness of the NIT	.96	.94	.94
Mean effect size *d* (reference group 1933/36)		.50	.79
The difference of the test variance		F_1=.72; p>.05	F_2=.67; p>.05 F_3=1.01; p>.10

Note: The *g*-loadedness of the test is calculated using the formula reported by Jensen (1998). The differences are computed as *F* ratios and tested for significance with ($N_{\text{first group}-1}$; $N_{\text{second group}-1}$) (see Jensen, 2003). Subscript 1 means the comparison cohorts 1933/36 and 1997/98; subscript 2 means the comparison cohorts 1933/36 and 2006; subscript 3 means the comparison 1997/98 and 2006.

Tables 2 and 3 show that between 1997/98 and 2006 scores on subtests A3 (Concept Comprehension) and B1 were decreasing, so these two tests strongly influence recent outcomes. If we exclude their results from the analysis, recent gains on the remaining eight tests are huge: approximately 5 IQ points in 8 years. So, the data show that the FE continues in Estonia into the present day and even show an acceleration of the effect.

Eight different grade groups took ten subtests each, resulting in 80 different data sets. Use of the Kolmogorov–Smirnov test showed that, in general, the score distributions deviated somewhat from normality. Most of the subtests have negative skewness and kurtosis. The 8th grade sample from 1997/98 showed the largest deviation from normality on the subtest A4 (Synonyms–Antonyms): skewness=−1.62 and kurtosis=3.60. Comparable values are commonly found in the large majority of test batteries.

3.2. The *g* factor and the *g*-loadedness of NIT in different decades

We analyzed the *g* factors using two techniques. First, we compared factor solutions on the basis of factor loadings to ensure comparability with previous FE research. Secondly, we compared the *g* factors using Multigroup Confirmatory Analysis (MGCFA) to provide a more comprehensive overview of comparability of the *g* factors in different periods for different age groups. The *g* factors were extracted using one-factor (all 10 subtests are regressed to the common latent *g*) confirmatory analysis (CFA) of LISREL (Jöreskog & Sörbom, 2006). The comparison of factor loadings using the *F* statistic (see Table 4) shows that the *g* factor of the cohort from 1933/36 is significantly different from that of later cohorts (from 1997/98 and 2006), but the *g* factors of the two most recent groups are highly similar. The *g*-loadedness of the NIT sum scores in 1934/36, 1997/98, and 2006 was calculated using the

Table 5
Equality constraints of model parameters

Model	Factor loadings	Residual variances	Intercepts	Factor means
1 (configural invariance)	Free	Free	Free	Fixed at 0
2 (metric invariance)	Invariant	Free	Free	Fixed at 0
3 (equal residuals)	Invariant	Invariant	Free	Fixed at 0
4 (strict invariance)	Invariant	Invariant	Invariant	Free

Note. Each model is nested under the previous one.

Table 6
Fit Indices test for factorial invariance of NIT 1997/98–2006

Model	Equality constraints	χ^2	*df*	Compare	$\Delta\chi^2$	Δdf	RMSEA	CFI	AIC	CAIC
Grade 7										
1	–	98.5	70				.038	.993	218	539
2	loadings	107.0	79	1 vs. 2	8.5	9	.035	.993	209	481
3	+residuals	105.2	89	2 vs. 3	1.8	10	.025	.996	187	406
4	+intercepts	196.6	98	3 vs. 4	91.4	9	.060	.975	261	431
4a	−except: A1, A3, B1	120.1	95	4a vs. 4	76.5	3	.031	.994	190	377
Grade 8										
1	–	79.7	70				.024	.997	200	508
2	loadings	90.8	79	1 vs. 2	11.1	9	.025	.996	193	455
3	+residuals	85.2	89	2 vs. 3	5.6	10	.000	1.000	167	378
4	+intercepts	156.9	98	4 vs. 3	71.7	9	.051	.979	220	386
4a	−except: A3, A4	110.7	96	4a vs. 4	46.2	3	.026	.995	179	353

Note. The Satorra–Bentler χ^2 statistic is used. A letter after model number indicates that the basic model is changed via freeing the parameter of the subtest (second column) responsible for deterioration in model fit.

formula given by Jensen (1998, pp. 103/104). Table 4 shows that although the *g*-loadedness of the composite score dropped from .96 in 1934 to .94 in 2006, the NIT in both cases is highly, almost perfectly, *g*-loaded. The factor loadings in this table are presented in standardized form to ease interpretation, but the analyses are based on the covariance matrices. The comparability of the *g* factors was further scrutinized by comparing different periods for different age groups.

3.3. Comparability of one-factor solutions

CFA yields one-factor solutions that fit all three samples well. The chi-square test does not support the one-factor models, but the fit indices that do not depend on the size of the sample indicate a good fit: the Root Mean Square Errors of Approximation (RMSEA) are, respectively, .029, .053, and .038; Non-Normed Fit Index (NNFI) is, respectively, 1.00, .98 and .99; Comparative Fit Index (CFI) is, respectively, 1.00, .99 and .99.

Although the one-factor models hold in all samples from the 1930s as well as more recent data, the regression of subtest scores on *g* was not invariant in an earlier exploratory analysis of Estonian NIT data (Wicherts et al., 2004). The preliminary analysis of the present data yielded mainly the same result: only configural measurement invariance in a comparison of the cohorts from 1933/36 and 2006 and weak measurement invariance in a comparison of the cohorts from 1997/98 and 2006. A possible cause of the lack of strict measurement invariance is age/grade heterogeneity of samples.

In this study we created homogeneous samples. We divided samples into subsamples of age/grade groups so as to compare three pairs of age/grade groups with MGCFA in comparison with the samples from 1933/36 and 2006 and two pairs age/grade groups in comparison with the samples from 1997/98 and 2006. Several estimations of similarity of the regression parameters allow us to be more confident that the estimations of differences do not depend significantly on specificity of samples or testing situations. For example, comparing the differences of the *g* factor models 1933/36 and 2006 we have three different estimations of all regression parameters. If all three comparisons yield the same result, then we can be confident in our conclusions about differences in the *g* factor models. For describing the *g* factor solutions from period 1997/98 and 2006 we have only two age groups and factor solutions respectively, but the logic of the conclusions is the same.

In the current paper we followed the common approaches to estimate the comparability of latent variables (see Meredith, 1993; Widaman & Reise, 1997, Widaman & Thomson, 2003; Meredith & Teresi, 2006). Measurement invariance was investigated using the strategy employed by Wicherts et al. (2004) and Wicherts and Dolan (submitted for publication) as this modeling logic was used in the FE research previously. The measurement invariance is estimated by fitting a series of increasingly restrictive models of covariance and means of subtests simultaneously (Table 5). At first the model with no restriction is imposed (configural invariance). Step 2: the factor loadings are restricted (metric invariance). Step 3: the factor loadings and residual variances are restricted (equal residuals). Step 4: the mean structure is investigated by restricting intercepts, factor loadings and residual variances, and freeing factor means (strict invariance). The restrictions in step 4 create a model where observed mean differences are due to common factor mean differences. This estimation is important for understanding secular changes in test scores. It is possible that the FE is mainly caused by the familiarity of different cohorts with item types. Without strict factorial invariance one is not able to draw clear conclusions concerning group differences.[1]

The fit of MGCFA models is assessed by the chi-squared statistics in relation to degrees of freedom and other fit indices: RMSEA, CFI, Akaike's Information Criterion (AIC), and Consistent Akaike's Information Criterion (CAIC). The stepwise approach is used, in which increasingly more constraints are introduced after which the model fit is assessed (Tables 6–7). If a given constraint leads to a clear deterioration in fit, the source responsible for misfit is found via modification indices and a new model without non-invariant parameter is assessed. The improvement of modeling at each restriction step is stopped if the next changes do not significantly improve the initial model (in chi-square terms) or the next change in the model is not more parsimonious than the previous one (AIC or CAIC is higher than in previous model).

[1] Strong factorial invariance is less restrictive because it does not include the equality constraint on the residual variances. Strong factorial invariance was tested but the results did not provide additional information as compared to the results with the most restrictive model.

Table 7
Fit indices test for factorial invariance of NIT 1933/36–2006

Model	Equality constraints	χ^2	*df*	Compare	$\Delta\chi^2$	Δdf	RMSEA	CFI	AIC	CAIC
Grades 4/6 (the youngest group)										
1	–	91.4	70	–	–	–	.035	.993	211	526
2	loadings	137.3	79	2 vs. 1	45.9	9	.054	.982	239	507
2a	−except: A4, B1, B2	105.0	76	2a vs. 2	32.3	3	.039	.991	213	496
3	+residuals	140.7	86	3 vs. 2a	35.7	10	.050	.983	229	459
3a	−except: A1	124.8	85	3a vs. 3	15.9	1	.043	.988	215	451
4	+intercepts	303.0	91	4 vs. 3a	178.2	6	.096	.935	381	585
4a	−except: A5, B3, B5	124.8	88	4a vs. 4	178.2	3	.040	.989	209	429
Grades 5/7 (the middle group)										
1	–	73.4	70				.013	.999	193	514
2	loadings	101.1	79	2 vs. 1	27.6	9	.031	.995	203	475
3	+residuals	129.7	89	3 vs.2a	28.6	10	.040	.991	212	430
3a	−except: A4	103.0	88	3a vs.3	26.7	1	.025	.997	187	411
4	+intercepts	524.3	96	4 vs.3a	421.3	8	.126	.904	592	774
4a	−except: A1, A5, B1, B2, B3, B5	105.9	90	4a vs.4	418.4	6	.025	.996	186	399
Grades 6/8 (the oldest group)										
1	–	96.8	70	–	–	–	.032	.994	217	553
2	loadings	109.5	79	2 vs. 1	12.7	9	.032	.993	211	497
3	residuals	116.2	89	3 vs. 2	6.7	10	.029	.994	198	428
4	intercepts	690.6	98	4 vs. 3	574.4	9	.129	.865	754	933
4a	Except: A1, A5, B1, B2, B5	132.2	93	4a vs. 4	558.4	5	.034	.991	206	413

Note. The Satorra–Bentler χ^2 statistic is used. A letter after model number indicates, that the basic model is changed via freeing the parameter of the subtest (second column) responsible for deterioration in model fit.

In principle, all subtests of NIT (8 subtests from 10) may cause differences in regression parameters. At the same time, invariant model parameters may be different for different group comparisons. Therefore, the invariance evaluations of regression parameters were made for different group comparisons and the regression parameters were scrutinized for subtests (see Tables 6–7). The combined results allow us to draw more generalized conclusions.

3.4. Differences in factor loadings and regression intercepts

There is no unequivocal evidence from different comparisons about differences in factor loadings. In all comparisons the factor loadings are invariant with the exception of the youngest groups from 1933/36 and 2006. Fit indices of the last initial model (model 2) assuming equality in factor loadings were relatively good (RMSEA=.054; CFI=.982), and freeing factor loadings of subtests A4, B1, and B2 improved the values of the model fit indices a little (model 2a). Generally the factor loadings are invariant in our comparisons.

The main reason of the absence of strict measurement invariance of the measurement is the difference in measurement intercepts. This means that the students at the same level of mental ability (constant latent *g*) showed different results in NIT subtests (observed subtest scores).

3.5. Comparison of the intercepts of the regression models of 1997/98 and 2006

Clearly the subtest A3 (Concept Comprehension) has different intercepts in the regression models, as its intercepts were not invariant in either comparisons (Table 6). The intercepts of subtests A1 (Arithmetic), A4 (Synonyms–Antonyms), and B1 (Computation) were non-invariant in only one comparison. There were unexpected decreases in test scores in subtests A3 and B1 also in comparison with recent data (Tables 2 and 3). Evidently the regression intercepts of these subtests are not invariant. Excluding the results of those two subtests from the analysis yields an IQ gain of approximately .30 SD for a period of eight years.

3.6. Comparison of the intercepts of regression models of 1933/36 and 2006

Clearly the *g* models of 1933/36 and 2006 differ by regression intercepts (Table 7). In all three comparisons the subtests A5 (Symbol–Number) and B5 (Comparisons) have different intercepts. In two comparisons from three subtests A1 (Arithmetic), B1 (Computation), B2 (Information), and B3 (Vocabulary) regression intercepts were not invariant. It is evident that in 2006 the subtest A5 and B5 do not have the same meaning they had in 1933/36. The comparison of the cohorts on the bases of those subtests will give "hollow" results. The conclusions about gains based on the subtest A1, B1, B2, and B3 should also be made with caution.

In the initial stage (model 4), models testing the equality of intercepts yielded bad fit estimations. Table 7 shows, for instance, that comparing data from1933/36 and 2006 using data from older children yielded values of RMSEA=.129, and CFI=.865. Thus, it can be concluded, that when comparing the data from 1933/36 and 2006 there are some minimal differences in factor loadings, but the main and significant differences are in regression intercepts. This means, first of all, that students at the same level of general mental ability (*g*) from different cohorts have different manifest test scores: *g* has different impact on the performance of students in different subtests in different cohorts making some subtests clearly easier for later cohorts.

4. Discussion

We analyzed old and recent data on the Estonian National Intelligence Test (NIT) with 72 years in between; this is the largest time interval in the FE literature. This study provides evidence that the FE in Estonia continues and that the meaning of test-score changes over time.

4.1. IQ scores are still going up in Estonia

The Estonian data in this study do not show the standstill in IQ scores as clearly shown in the more highly developed Scandinavian countries for quite a few years or the decrease in IQ scores of recent years (Sundet, Borren, & Tambs, 2007; Teasdale & Owen, 2007). Actually, the IQ gain per decade is higher in younger student groups than in older ones when comparing the samples 1933/36 and 2006. But the recent (1997/98–2006) gains are almost the same for different age groups — about 1/6 SD in eight years. There is no tendency for the gains to diminish with age in the most recent period.

The gains differ by subtests and scores on two subtests – B1 (Computation), A3 (Concept Comprehension) – have decreased. Leaving out these two subtests yields an IQ increase of about .30 SD, or approximately 5 IQ points in only eight years. This means that the gain of present day Estonian students is twice the typical gain of 3 IQ points per decade.

The decreasing scores on Computation go with increasing scores on Arithmetic: the discrepancy is 8.5 IQ points in the last 8 years. Although the subtests' names are similar, reflecting the fact that the test constructors meant them to be highly similar, their content is different. Arithmetic subtest consists of mathematical questions requiring reasoning, but Computation requires typical calculus operations; the skills necessary for the Arithmetic subtest are needed in contemporary everyday life, but the manual computation skills are needed only rarely.

The rise in Arithmetic and in most language-based subtests scores is different from Flynn's (2007) results: In the US, Britain and elsewhere, the subtests of the WISC conspicuous for nil/low gains are Arithmetic (A1), Information (B2), and Vocabulary (B3). In Estonia the gains on these three subtests are huge in recent years, averaging at .36 SD or 5.4 IQ points in only eight years. Flynn (2007) reports large US gains for WISC Coding and WISC Similarities, and we see the same pattern for, respectively, NIT subtests A5 and B5. The NIT's subtest Computation (B1) does not have an equivalent in the WISC, but the findings are compatible with Flynn's (2007) analysis: manual computation is not an everyday skill in modern society, so one would expect a decrease in scores. So, the Estonian gains show a quite strong compatibility with Flynn's (2007) analysis.

Flynn (2007) hypothesizes social multipliers, including education, have led to people thinking more scientifically and logically, resulting in higher IQ scores. The large, recent gains in Estonia might be explained the same way. After becoming an independent country again, it faced profound reforms in education. In principle a new curriculum of general education was applied. Krull and Trasberg (2006) gave as keywords for the new curriculum: learner-centered, integration of different subjects, focused on general competencies in terms of expected learning outcomes, social and communication skills and values. The improvement of the educational system is most likely reflected in the excellent results of Estonian students in the fields of mathematics and science in international comparison. For example, in the third international mathematics and science study (TIMSS) in 2003 Estonian 8th graders had one of the highest achievement scores (Mullis, Martin, Gonzalez & Chrostowski, 2003; Martin, Mullis, Gonzalez & Chrostowski, 2003). Finally, other main reasons for test-score rise are likely to be better nutrition, better health care, and changes in demographical behavior.

4.2. The relation between gains and g

The relation between *g* and tests score gains is central to FE research. Where Flynn (2007) argues that there is no gain at all on *g*, the gains are all on broad and narrow abilities, Jensen (1998) suggests that less than half of the FE constitutes a real gain on *g*, but the other half are "hollow" gains. However, the use of the term "hollow" does not mean that the secular changes have no social significance. Six NIT subtests have clearly different meaning in different periods. The fact that the subtest Information (B2) has got more difficult may signal the transition from a rural to an urban society. Agriculture, rural life, historical events and technical problems were common in the 1930s, such as items about the breed of cows or possibilities of using spiral springs, whereas at the beginning of the 21st century students have little systematic knowledge of pre-industrial society. The fact that tasks of finding synonyms–antonyms to words (A4) is easier in 2006 than in the 1930s may result from the fact that the modern mind sees new choices and alternatives in language and verbal expression. More clearly the influence of language changes was revealed in several problems related to fulfilling subtest A4 (Synonyms–Antonyms). In several cases contemporary people see more than one correct answer concerning content and similarities or differences between concepts. It is important that in his monograph Tork (1940) did not mention any problems with understanding the items. It seems that language and word connotations have changed over time.

The sharp improvement in employing symbol–number correspondence (A5) and symbol comparisons (B5) may signal the coming of the computer game era. The worse results in manual calculation (B1) may be the reflection of calculators coming in everyday use.

4.3. Comparability of IQ scores over time

Analysis of measurement invariance significantly contributes to understanding why a big part of IQ test gains is "hollow". In the FE structural modeling two outcomes are highly informative: the estimation of invariance of factor loadings and of regression intercepts. We have not found univocal evidence for the lack of invariance in factor loadings in group comparisons at the subsamples level. Only in one comparison out of five the factor loadings were different. However, some of the regression intercepts were different in all model comparisons. The absence of strict measurement invariance between cohorts leads to the general conclusion that IQ scores change their meaning over time.

The lack of invariance across groups has several possible causes, all of which imply that the model is not a proper one for the data. One problem may be the set of tests, which do not seem to conform to Spearman's stipulations when designing a battery of tests to identify a general factor. Spearman sought a battery of tests without too much overlap among tests. At present, several of the subtests seem to have similar content,

and this differs from that of other subtests. The differential overlap of test content almost surely means that a single-factor model is not appropriate for the data; invariance tests of one-factor models did not fare well. This lack of invariance has clear implications for estimating overall general intelligence. With lack of invariance of the *g* factor, overall statements about Flynn Effects on general intelligence are unjustified.

5. Conclusion

Our analyses reveal that the secular gain in IQ test scores continues in Estonia, and that their meaning has changed over time. Most of the subtests are now easier and more familiar to students than six or seven decades ago. When the subtests scores are regressed unto the underlying *g* factor not all regression parameters are invariant for different cohorts. The differences in regression intercepts are the main cause of the lack of measurement invariance. So, IQ scores in 1933/36, 1997/98, and 2006 clearly do not have the same meaning, making a direct comparison of mean IQ scores of various cohorts impossible. A comparison of the subtests scores of cohorts at the level of concrete activities and skills over time leads to various fruitful explanations of secular gains.

Acknowledgment

The authors express their appreciation to Jelte Wicherts, James Flynn, Conor Dolan and to an anonymous reviewer for their help in preparing the manuscript.

Appendix A

Correlations of subtests

	A1	A2	A3	A4	A5	B1	B2	B3	B4	B5
4th grade in 1933/36 (above the diagonal, *n*=270), 6th grade in 2006 (below the diagonal, *n*=243)										
A1		.510	.461	.424	.143	.151	.457	.319	.457	.002
A2	.410		.491	.496	.155	.164	.577	.390	.524	.185
A3	.374	.484		.504	.290	.100	.577	.386	.504	.289
A4	.395	.538	.360		.241	.074	.618	.432	.497	.226
A5	.254	.224	.302	.292		.123	.192	.098	.279	.270
B1	.497	.384	.303	.341	.275		.193	.143	.270	.055
B2	.437	.459	.496	.468	.257	.416		.557	.634	.261
B3	.368	.474	.433	.425	.221	.302	.549		.470	.234
B4	.463	.569	.581	.445	.311	.400	.534	.462		.257
B5	.236	.255	.227	.235	.300	.342	.239	.241	.374	
5th grade in 1933/36 (above the diagonal, *n*=222), 7th grade in 2006 (below the diagonal, *n*=343)										
A1		.530	.500	.408	.246	.413	.477	.361	.494	.260
A2	.417		.657	.503	.231	.366	.648	.582	.621	.357
A3	.358	.508		.473	.256	.304	.690	.543	.586	.330
A4	.402	.490	.407		.268	.229	.555	.503	.479	.364
A5	.255	.272	.305	.325		.283	.300	.234	.298	.342
B1	.547	.507	.350	.382	.365		.349	.267	.401	.327
B2	.489	.591	.468	.471	.339	.423		.615	.596	.412
B3	.333	.440	.394	.368	.211	.303	.504		.497	.321
B4	.469	.535	.519	.417	.349	.445	.536	.472		.385
B5	.288	.395	.254	.349	.468	.450	.329	.223	.376	
6th grade in 1933/36 (above the diagonal, *n*=407), 8th grade in 2006(below the diagonal, *n*=327)										

Appendix A (*continued*)

	A1	A2	A3	A4	A5	B1	B2	B3	B4	B5
A1		.443	.398	.404	.185	.395	.496	.332	.419	.145
A2	.392		.526	.425	.222	.262	.537	.385	.426	.194
A3	.415	.533		.456	.219	.204	.524	.479	.390	.185
A4	.356	.496	.497		.242	.201	.491	.348	.356	.320
A5	.192	.327	.264	.364		.247	.208	.153	.316	.326
B1	.461	.325	.329	.320	.316		.271	.202	.275	.246
B2	.493	.546	.495	.487	.176	.373		.525	.427	.298
B3	.331	.437	.421	.434	.148	.234	.492		.291	.126
B4	.423	.535	.495	.487	.347	.350	.495	.452		.280
B5	.214	.355	.289	.308	.456	.335	.248	.148	.356	
7th grade in 1997/98 (above the diagonal, *n*=246), 8th grade in 1997/98 (below the diagonal, *n*=164)										
A1		.458	.243	.288	.116	.411	.497	.406	.450	.185
A2	.388		.474	.433	.238	.356	.617	.545	.515	.301
A3	.085	.384		.402	.265	.296	.503	.392	.454	.314
A4	.355	.428	.279		.290	.357	.400	.390	.465	.279
A5	.226	.154	.210	.123		.295	.159	.132	.178	.421
B1	.384	.315	.140	.236	.309		.372	.297	.385	.428
B2	.440	.571	.347	.381	.261	.279		.558	.563	.349
B3	.393	.479	.259	.389	.164	.267	.561		.438	.260
B4	.397	.487	.246	.411	.324	.433	.512	.385		.256
B5	.269	.142	.224	.099	.356	.306	.197	.179	.233	

References

Ceci, S. (1991). How much does schooling influence general intelligence and its cognitive components? A reassessment of the evidence. *Developmenta Psychology*, *27*, 703–722.

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

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

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

Dickens, W., & Flynn, J. (2002). The IQ paradox is still resolved: Reply to Loehlin (2002) and Rowe and Rodgers (2002). *Psychological Review*, *109* (4), 764–771.

Emanuelsson, I., & Svensson, A. (1986). Does the level of intelligence decrease? A comparison between thirteen years-olds tested in 1961, 1966 and 1980. *Scandinavian Journal of Educational Research*, *30*, 25–38.

Emanuelsson, I., Reutenberg, S., & Svensson, A. (1993). Changing difference in intelligence? Comparison between groups of 13-year-olds tested from 1960 to 1990. *Scandinavian Journal of Educational Research*, *3*, 259–277.

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

Flynn, J. R. (2000). IQ gains, WISC subtests, and fluid *g*: *g* theory and the relevance of Spearman's hypothesis to race (followed by discussion). In G. R. Bock, J. A. Goode, & K. Webb (Eds.), *The nature of intelligenceNovartis Foundation Symposium*, *vol. 233*. (pp. 202–227) New York: Wiley.

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

Jensen, A. R. (1998). *The g factor*. Westport, CT: Praeger.

Jensen, A. R. (2003). Regularities in Spearman's law of diminishing returns. *Intelligence*, *31*, 95–105.

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

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

Jöreskog, K., & Sörbom, D. (2006). *LISREL 8.80*. Lincolnwood, IL: Scientific Software International.

Krull, E. & Trasberg, K. (2006). Changes in general education from collapse of the Soviet Union to EU entry. ERIC on-line submission. (Eric Document reproduction Service No. ED495353).

Lynn, R. (1990). The role of nutrition in secular increases in intelligence. *Personality and Individual Differences*, *11*, 273–285.

Lynn, R. (1998). In support of the nutrition theory. In U. Neisser (Ed.), *The rising curve. Long-term gains in IQ related measures* Washington: American Psychological Association.

Lynn, R., & Harvey, J. (2007). The decline of the world's IQ. *Intelligence*. doi:10.1016/j.intell.2007.03.004

Meisenberg, G., Lawless, E., Lambert, E., & Newton, A. (2006). The social ecology of intelligence on a Caribbean Island. *Mankind Quarterly*, *46*, 395–433.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrica*, *58*, 523–543.

Meredith, W., & Teresi, J. (2006). An essay on measurement and factorial invariance. *Medical Care*, *44*, S69–S77.

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

Mingroni, M. (2007). Resolving the IQ paradox: Heterosis as a cause of the Flynn effect and other trends. *Psychological Review*, *114*, 806–829.

Martin, M., Mullis, I., Conzalez, E., & Chrostowski, S. (2003). *TIMSS 2003. International science report.* Boston: TIMSS & PIRLS International Study Center.

Mullis, I., Martin, M., Conzalez, E., & Chrostowski, S. (2003). *TIMSS 2003. International mathematics report.* Boston: TIMSS & PIRLS International Study Center.

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

Neisser, U. (1998). Introduction: Rising test scores and what they mean. In U. Neisser (Ed.), *The rising curve. Long-term gains in IQ and related measures* (pp. 3–22). American Psychological Association.

Shayer, M., Ginsburg, D., & Coe, R. (2007). Thirty years on — A large anti-Flynn effect? The Piagetian test Volume & Heaviness norms 1975–2003. *British Journal of Educational Psychology*, *77*, 25–71.

Statistics Estonia (2007). *Pupils in full-time general education by year, county, grade and language of instruction.* [Statistical Database], [Retrieved June, 28, 2007, from http:// www.stat.ee].

Sundet, J., Borren, I., & Tambs, K. (2007). The Flynn effect is partly caused by changing fertility patterns. *Intelligence*. doi:10.1016/j.intell.2007.04.002

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

Teasdale, T., & Owen, D. (1987). National secular trends in intelligence and education: A twenty-year cross-sectional study. *Nature*, *325*, 119–121.

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

Teasdale, T., & Owen, D. (2007). Secular declines in cognitive test scores: A reversal of the Flynn Effect. *Intelligence*. doi:10.1016/j.intell.2007.01.007

Terman, L. M. (1921). *The intelligence of school children.* London: George G. Harrap.

Tork, J. (1940). *Eesti laste intelligents [The intelligence of Estonian children.] Estonia, Tartu: Koolivara.*

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

Wicherts, J., & Dolan, C., (submitted for publication). Measurement invariance and group differences in intercepts in confirmatory factor analysis.

Widaman, K., & Thomson, J. (2003). On specifying null model for incremental fit indices in structural equation modelling. *Psychological Methods*, *8*, 16–37.

Widaman, K., & Reise, S. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* Washington: American Psychological Association.