Intelligence 41 (2013) 770–779

Contents lists available at ScienceDirect

Intelligence

An item-level examination of the Flynn effect on the National Intelligence Test in Estonia

William Shiu [a], A. Alexander Beaujean [a,*], Olev Must [b], Jan te Nijenhuis [c], Aasa Must [d]

[a] Baylor University, United States
[b] Tartu University, Estonia
[c] University of Amsterdam, The Netherlands
[d] Estonian Defence College, Tartu, Estonia

ARTICLE INFO

Article history:
Received 9 December 2012
Received in revised form 6 May 2013
Accepted 26 May 2013
Available online 20 June 2013

Keywords:
Flynn effect
Item response theory
Estonia
Invariance

ABSTRACT

This study examined the Flynn effect (FE; i.e., the rise in IQ scores over time) in Estonia using the Estonian version of the National Intelligence Tests (NIT; Haggerty, Terman, Thorndike, Whipple & Yerkes, 1919; National Research Council, 1920). Using secondary data from two cohorts (1934, $n = 890$ and 2006, $n = 913$) of students, we analyzed the NIT's subtests using item response theory (IRT). For each subtest, we first examined invariance in all the items and then linked the latent variable (θ) scores between the two cohorts using the invariant items. The results showed that there was a FE in θ for all subtests except one, although there was much variability in the FE magnitude, ranging from an effect size of 0.24 (3.60 IQ points) to 1.05 (15.75 IQ points). In addition, this study showed there was a decrease in the variability of θ for all the subtests, although only two of the subtests showed large decreases (approximately .50 standard deviations). Last, the subtests' precision of measuring θ was very similar at both time points.

1. Introduction

The Flynn effect (FE) is the rise in IQ scores over time (approximately 3 IQ points per decade or .3 points per year; Flynn, 2007; Neisser, 1998). The FE has been found on every inhabitable continent (e.g., Flynn, 1987; Flynn & Rossi-Casé, 2012; Pietschnig, Voracek, & Formann, 2010; te Nijenhuis, Cho, Murphy, & Lee, 2012; te Nijenhuis, Murphy, & van Eeden, 2011), and across a wide range of abilities (e.g., Howell, 2008; Kanaya, Scullin, & Ceci, 2003; Sanborn, Truscott, Phelps, & McDougal, 2003; Wai & Putallaz, 2012). Moreover, due to its prominence, it is now a part of many legislative discussions that concern cognitive ability assessment (e.g., Ceci, Scullin, & Kanaya, 2003; Flynn, 2006; Kanaya & Ceci, 2007; Young, Boccaccini, Conroy, & Lawson, 2007).

Although many causal theories have been put forth, the cause of the FE remains inconclusive. Some think the FE represents a genuine rise in cognitive ability due to, e.g., better nutrition (Cohen, Flament, Dubos, & Basquin, 1999; Lynn, 2009; Sigman & Whaley, 1998), increased cognitive stimulation (Blair, Gamsonb, Thornec, & Bakerd, 2005; Teasdale & Owen, 1987), or change in family structure and fertility patterns (Mingroni, 2007). Other researchers argue that the FE could just as easily be due to changes in the test as it is due to changes in the individuals taking the test (i.e., psychometric artifact; Beaujean & Osterlind, 2008; Brand, 1987), and argue that more psychometric work in the FE needs to be done before any strong causal theories should be developed (McGrew, 2010; Rodgers, 1998).

1.1. Measuring the Flynn effect

FE studies typically involve one of two types of designs. The first design is to examine scores from two or more versions of the same instrument (or two different instruments

* Corresponding author at: Department of Educational Psychology, One Bear Place #97301, Waco, TX 76798-7301, United States. Tel.: +1 254 710 1548; fax: +1 254 710 3265.
E-mail address: Alex_Beaujean@Baylor.edu (A.A. Beaujean).

0160-2896/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.intell.2013.05.007

normed at different times) administered to a single sample at a single time point (e.g., Covin, 1977). The second design to assess the FE is to compare data from a single instrument administered to two or more (ostensibly) similar samples from different generations (e.g., Flynn, 1987). For either design, the most common method used to measure the FE is to compare mean differences in aggregated scores (e.g., Full scale IQ, Verbal IQ). In doing so, investigators make an implicit assumption that the test scores are measuring the sameconstruct(s) the same way (i.e., invariance; Meredith & Teresi, 2006; Millsap, 2011).

1.2. Measurement invariance

The validity of between-group test score comparisons is threatened if items operate differently among groups (Kane, 2006; Messick, 1989; Ployhart & Vandenberg, 2010). If one cannot be sure that an instrument is measuring the same construct the same way at both time periods, then one cannot be sure that any observed group difference of the instrument's score are due to measuring the constructs differently or a true difference between the groups (Little, 1997; Steenkamp & Baumgartner, 1998; Thompson & Green, 2006; Vandenberg & Lance, 2000). Beaujean and Sheng (2013) make the following analogy comparing means from non-invariant test scores is akin to comparing average temperatures at two different geographic locations with thermometers that use different scales. While mean differences could be due to different temperatures, they could also be the result of the scales having different origins (e.g., Fahrenheit vs. Rankine), different units (e.g., Kelvin vs. Rankine), or both (e.g., Fahrenheit vs. Kelvin). Consequently, before comparing scores between groups, it is important to assess that the instruments are measuring the same construct, the same way (Millsap, 2011; Yoo, 2002).

1.2.1. Investigating measurement invariance

Typically invariance is examined using multi-group confirmatory factor analysis (MG-CFA) (Horn & McArdle, 1992). CFA is a very general latent variable framework that can handle both continuous indicators (traditional factor analysis) and categorical indicators (sometimes called binary factor analysis or item response theory [IRT]) (Bartholomew, Knott, & Moustaki, 2011). If there is at least strong invariance on the instrument among the groups, then group differences on the instrument's scores are due to group differences in the latent constructs the instrument is measuring and not a result of measurement artifacts or cultural differences. For the latent variables to be comparable among groups, at least three conditions must exist (Cheung & Rensvold, 2002; Little, 1997):

1. The indicators for the latent variable have the same configuration among the groups; that is, the groups should have the same number of latent variables, the same number of indicators, and the same pattern of fixed and free parameters (*configural invariance*).
2. The relationships between factors and indicators (i.e., loadings/pattern coefficients) are the same among the groups, thus establishing equivalence of the metrics of the latent variable(s) among groups (*weak/scalar invariance*).
3. The indicators' intercepts are the same among the groups, thus establishing equivalence of the latent variable's origin (*strong/scalar invariance*). In some situations, there is invariance in some indicators' loadings and intercepts, but not all of them, a situation Byrne, Shavelson, and Muthen (1989) called *partial invariance*. Although much more work needs to be done in this area (Vandenberg, 2002), the basic premise behind partial invariance is that as long as there is configural invariance and there are "enough" invariant indicators, the latent variable(s) can still be compared across groups. In such situations, the measures can be considered to be alternate forms: measuring the same latent variable, but using different indicators (although in situations involving the FE, there will be item overlap between the forms). To be able to compare the scores from the alternate forms, however, they must be first be equated.

1.2.2. Equating

The purpose of equating is to convert item and ability estimates from different measurement instruments (or alternate versions of the same instrument given to different populations) to a common scale to be able to compare the examinees' abilities (Baker, 1984; Dorans, 2004). There are two types of equating: horizontal and vertical. Horizontal equating is typically used to equate scores on alternate forms of a test given to equivalent groups, while vertical equating is typically used to equate scores on tests that differ in (overall) difficulty that are given to groups that differ in amount of the trait the test is measuring. For vertical equating, typically a common set of items (i.e., anchor items) is used across at least two versions of the test, which are used to determine the link needed to place the test scores on the same scale (Baker, 1984). With horizontal equating, the equivalence of the examinees allows for the alternate forms to be linked, although the equating process is greatly strengthened when there are anchor items used in this form of equating as well (Kolen & Brennan, 2004).

Such methods can readily be transferred to investigating the FE. If, as some hypothesize, subsequent generations are increasing in cognitive ability, then vertical equating can be used to transform the ability scales from different instruments for the different groups. On the other hand, if just the test properties are changing over time, then horizontal equating methods can be used to equate the scales from different measures for the equivalent groups.

1.3. Estonia

Estonia is a small country located in north-eastern Europe. The country covers approximately 45,200 km^2 and has a population of approximately 1.32 million. For many centuries, Estonia was the border between the western and the eastern world. The Estonian language belongs to the Finno-Ugrian branch of the Uralic language family, and many Estonians consider themselves to be a member of Nordic nations. Estonians declared their political independence in 1918, but the country was occupied by the Soviet Union in 1939/1940. After the collapse of Soviet regime in 1991, Estonia re-established its independence.

1.3.1. Previous investigations of cognitive ability in Estonia

Systematic investigations of cognitive abilities started in 1930s. One of the major contributions at this time was Tork's (1940) adaptation of the National Intelligence Tests (Haggerty et al., 1919) and his subsequent expatiation of the Estonian national IQ norms for schoolchildren. World War II and the Soviet occupation regulated psychological research, and suppressed intelligence research for decades. Although the soviet psychology never officially recognized individual differences in cognitive abilities, toward the end of 1960s the studies of cognitive abilities emerged (see Must & Allik, 2011). At that time, researchers investigated several IQ tests to find an appropriate one for Estonians. One of the first attempts was to adapt the Amthauer's (1953) *Der Intelligenz-Struktur-Test*. In the early 1970s, scholars began using the Raven's (1958) *Standard Progressive Matrices* (SPM), with Toim (1976) being the first to describe the Estonian student sample (1972/1973) in terms of intelligence with Raven data.

More recently, Lynn, Allik, Pullman, and Laidra (2002) reported on the Estonian standardization of the SPM on 2689 students from approximately 12 to 18 years of age. The results showed that the average IQ was 100.2 for Estonia, similar to the British IQ of 100. Similar results were found by Pullmann, Allik, and Lynn (2004) and Lynn, Pullmann, and Allik (2003), when comparing scores from Estonian children to other European countries.

1.3.2. Previous work on the Flynn effect in Estonia

Must, Must, and Raudik (2003a) examined literacy scores of 522 Estonian school children who were 9 and 14 years old in 1999 and compared them to 1994 literacy scores. The results showed that the 1999 cohort performed better than the 1994 cohort, which they concluded was a true gain in literacy (due to environmental and educational factors), but not a gain in general intelligence (*g*; Jensen, 1998).

Must, Must, and Raudik (2003b) investigated the FE in Estonian school children from 1933 to 1997. These authors used archived data and recent data collected on the National Intelligence Test (NIT; Haggerty et al., 1919). The archived data was from Tork's (1940) dissertation work examining the mental abilities of children in Estonia. As part of his dissertation, Tork adapted the "Anglicised" version of the NIT (National Research Council, 1920, cf. Ballard, 1922) to Estonian, and gathered data on approximately 3000 children from 12 to 14 years in grades 5 and 6, mainly from the city of Tartu. Must et al. concluded that that there was an overall increase in scores from 1933 to 1997. The gains were especially noticeable in the Analogies, Sentence Completion, Symbol–Number, and Comparisons subtests, which did not have very high *g*-loadings. Upon further investigation, they found a negative correlation ($-.40$) between the magnitude of the FE and the rank order of the *g*-loadings.

Must, te Nijenhuis, Must, and van Vianen (2009) continued the Must et al. (2003b) line of study by examining the FE in Estonian school children on the NIT across three cohorts from 1933, 1997, and 2006. They found a FE from 1933 to 2006 of approximately 1.65 IQ points per decade. In addition, they examined invariance of the Estonian NIT, using the subtest scores as indicators and fitting a single-factor model. Using MG-CFA, They found "minimal differences in factor loadings" from 1933 to 2006, but found relatively large differences in the intercepts. They concluded that students at the same *g*-level 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" (p. 30).

1.4. Current study

The current study is a continuation of the work of Must et al. (2009) and Must et al. (2003b). Using the NIT (Haggerty et al., 1919), we examined invariance at the item-level across the individual subtests from 1933 to 2006. Subsequently, if at least partial measurement invariance was exhibited for the subtests across the time cohorts, we then examined if the subtests, individually, exhibited a Flynn effect.

2. Method

2.1. Sample

The data comes from two samples of Estonian school children. The first sample comes from students ($n = 899$) who were part of the original Estonian National Intelligence Test (Tork, 1940) standardization sample. These students were gathered in 1933/36 ($\bar{x}$: 13.4 years, σ: 1.31 years). The second sample was gathered in 2006 ($n = 913$) ($\bar{x}$: 13.5 years, σ: .93 years). The second sample came from the same region as the first sample (Must et al., 2009). For more information about the sample, see Must et al. (2009).

2.2. Instrument

The instrument used for this study is the Estonian version of the "Anglicised" National Intelligence Test (NIT; Haggerty et al., 1919; National Research Council, 1920). Due to the success of the Army Alpha/Beta Intelligence Exams (Yerkes, 1921), Haggerty et al. developed the NIT for the purpose of measuring cognitive ability in school children, applying the method of group intelligence examination used in the military during World War I (Yerkes, 1921). The goal of the original NIT was to create a diverse set of tests in a single booklet that could be administered to any child who could read, write and participate in a group examination (Whipple, 1921). The NIT then "found [its] way to England and [was] issued in an Anglicised form by Harrap & Co." (Ballard, 1922, pp. 6–7).

The NIT (Haggerty et al., 1919) is comprised of 10 subtests across two scales, Scale A and Scale B (each scale containing 5 subtests). Scale A consists of the Arithmetic, Sentence Completion, Synonyms–Antonyms, Symbol–Digit, and Logical Selections subtests, while Scale B consists of the Computation, Information, Vocabulary, Analogies, and Comparisons subtests. As the Symbol–Digit and Comparisons subtests were designed specifically to measure processing speed, they were not included in the current analysis as such tests typically require specialized models (e.g., van der Linden, 2007). The items of each of the eight subtests used for this study are arranged progressively from least difficult to most difficult (see Fig. 1 for sample items). Each NIT subtest is timed, taking 2–4 minutes. Consequently, all missing data in the current study were scored as incorrect responses. (For an

Subtest	Directions
Analogies	Read carefully the first three words in each line. Then read the last four and draw a line under the right one. Example Item: cats-meow—cows–tail bark moo barn milk
Arithmetic	Find all the answers as quickly as you can Example Item: How long is half of 8 minutes?
Computation	Do this work in arithmetic as quickly as you can without making mistakes. Example Item: Multiply 2 × 3 =
Information	In each sentence draw a line under the one word that makes the sentence true. Example Item: The number of quarters in a dollar is 2 3 4 5
Logical Selections	In each row draw a lint under each of the two words that tell what the thing always has. Example Item: man (head stick heart pants hair)
Sentence Completio	Write on each dotted line one word to make the sentence sound sensible and right. Example Item: Ice is cold, but fire is . . .
Synonyms-Antonyms	Write S if two words are the same; Write D if they are as different as can be. Example Item: dry . . . wet
Vocabulary	Read each question and draw a line under the right answer. Example Item: Are pears good to eat? Yes No

Fig. 1. Example items from the American version of the *National Intelligence Test*.

alternative approach to handling missing data with this dataset see Must and Must (2013-this issue).)

Tork (1940) adapted the American NIT to Estonia as part of his doctoral dissertation, norming it on over 6000 school children. In adapting the test, Tork (1940) made some modifications to every subtest except the Symbol–Digit subtest for his Estonian translation to include content specific to Estonia. Similar to the original NIT, the Estonian version comes with practice items on separate sheets. Test takers first complete the practice items as a group, and then complete the actual test items independently (Must et al., 2003b, 2009).

2.3. Data analysis

2.3.1. Latent variable model

For all subtests, we used the factor model shown in Eq. (1).

$$p\left(x_{ij}=1|\theta_j\right)=\frac{1}{\exp\left(a_i\left[b_i-\theta_j\right]\right)+1} \quad (1)$$

where

x_{ij} is the response of examinee j on item i,
a_i is the ith item's discrimination,
b_i is the ith item's difficulty,
θ_j is the jth examinee's ability on the (single) construct the test is measuring. In this model, $p(x_{ij}=1|\theta_j)$ is interpreted as the probability of examinee j, who has level θ_j on the latent variable θ, correctly answering item i, which has difficulty b_j and discrimination a_i. This model is equivalent to the more typical factor analytic model shown in Eq. (2).

$$p\left(x_{ij}=1|\theta_j\right)=f(\alpha_i\theta+\beta i) \quad (2)$$

where

α_i is the ith item's factor loading
β_i is the ith item's intercept, and
$f()$ is the cumulative logistic distribution (Bock & Moustaki, 2006). Of note, $a_i=\alpha_i$ and $b=\frac{-\beta}{\alpha}$.

2.3.2. Procedures

We examined the FE for each subtest using six steps.

1. Determine the dimensionality of the items for each subtest via exploratory factor analysis using the Normal Ogive Harmonic Analysis Robust Method (NOHARM; Fraser &

McDonald, 1988) program. NOHARM uses two indices to determine model fit: the goodness of fit index (GFI) and the root mean square residual (RMSR). GFI values closer to one indicate a better fitting model, with values of .90 or higher indicating the model fits the data relatively well (McDonald, 1999). The RMSR is an indicator of unconditional fit where a value of zero is a perfect fit, but values less than .08 are considered a good fit (Hu & Bentler, 1998).

2. Determine which item factor model best fits a given subtest's data: Rasch ($a = 1$ for all items), one-parameter (1P, single a value for all items, but $a \neq 1$), or two-parameter (2P, a_i and b_i allowed to vary across all items). We did this using Mplus (Muthen & Muthen, 2010) to fit 2P, 1P, and Rasch models to the item data for each subtest and comparing model fit between the models using the comparative fit index (CFI), Tucker–Lewis index (TLI), and root mean square error of approximation (RMSEA).
3. Test to see if the items exhibit invariance in both the discrimination (factor loading) and difficulty (intercept) between the 1933 and 2006 groups. We used multiple measures to determine if an item exhibited invariance: Mantel–Haenszel (MH; Holland & Thayer, 1988), logistic regression (Swaminathan & Rogers, 1990), transformed item difficulties (TID; Angoff & Ford, 1973), standardization (Std; Dorans & Kulick, 1983, 1986) and Breslow–Day (BD; Aguerri, Galibert, Attorresi, & Prieto Maranon, 2009; Penfield, 2003). For the purposes of this study, we defined an item to not exhibit invariance if at least three of the five statistics indicated invariance was not present. We did this using the difR (Magis, Beland, Tuerlinckx, & De Boeck, 2010) package in **R** (R Development Core Team, 2012).
4. Link the items exhibiting invariance, and then use the linked items to equate the 1933 and 2006 θ estimates. The items not exhibiting invariance were used to within a cohort to aid in estimating θ, whereas the anchor items were used to estimate θ as well as form the link needed to put the θ estimates from the two groups onto the same scale. We did this using the `ltm` (Rizopoulos, 2006) package in **R**.
5. Examine the θ scores between groups to see if there were any differences in the means, variances, or reliability. We measured reliability using the θ standard error (θSE; sometimes called the standard error of estimation), which is an information index. It gives a test's precision of measuring θ at a given level of θ and is calculated as the inverse of the square root of the test's information at a given level of θ (i.e., $\theta_j \mathrm{SE} = \frac{1}{\sqrt{I(\theta_j)}}$). The interpretation of θSE is comparable to the interpretation of "traditional" standard error of measurement because a smaller θSE indicates a higher reliability (Haertel, 2006).
6. Estimate effect sizes for the FE using traditional (classical test theory [CTT]) methods. We estimated cognitive ability as the sum score for all the items in a subtest.

3. Results

3.1. Dimensionality

Table 1 shows the results from extracting a single factor, and indicates that the subtests appear to be unidimensional.

3.2. Model fit

The 2P, 1P, and Rasch models were compared for data fit for all subtests. Based on the values of the CFI, TLI, and RMSEA the 2P model appears to fit the data better than the 1P and Rasch models for all the subtests. Consequently, it was the model used for the subsequent IRT-based analyses.

3.3. Item invariance

The number of items that did not meet the criteria for invariance is given in Table 2. Except for the Arithmetic subtest, where only 13% of the items did not show invariance, approximately one-third to one-half of the subtests' items did not show invariance. Consequently, we examined how much information (i.e., precision of the estimate of the latent variable; Hambleton & Swaminathan, 1985) is present in the invariant items for $-4 \leq \theta \leq 4$. We chose -4 and 4 as, if θ follows a normal distribution, this should cover the majority of the respondents. The test information curves for the invariant items for each subtest are given in Fig. 2, and indicate that most of information (from 74 to 97%) for the invariant items is in the $-4 \leq \theta \leq 4$ range. Moreover, the test information curves tend to be disbursed across a wide range of θ levels, so should work fairly well to anchor the subtests across groups.

3.4. Comparing scores across years

Table 3 displays the subtests' score average and variability in θ for the 1933 and 2006 groups. In addition, the table shows the IRT-version of reliability (mean θ standard error and the standard deviation of the θ standard error) for the two groups.

Instead of measuring if the difference in means was "statistically significant" across the two time periods, we calculated Hedges (1981) *g*. Like Cohen's (1988) *d*, Hedges' *g* is an effect size (ES) estimator measured in standard deviation units. To calculate Hedges' *g*, we subtracted the 1933 scores from the 2006 scores; thus, a positive ES indicates a gain over time (i.e., a "Flynn effect") and a negative number indicates a decrease over time (i.e., a "reverse Flynn effect"). The results show a FE for all subtests except Computation, where there was a reverse FE. The last

Table 1
Results from exploratory factor analysis using NOHARM program.

Subtest	RMSR	GFI
Analogies	0.01	0.96
Arithmetic	0.00	0.99
Computation	0.01	0.95
Information	0.01	0.91
Logical Selections	0.01	0.98
Sentence Completion	0.01	0.98
Synonyms–Antonyms	0.01	0.98
Vocabulary	0.01	0.91

Note. Results are for fitting a single factor model. RMSR: root mean square residual; GFI: goodness of fit index.

Table 2
Items not exhibiting invariance.

Subtest	Number of non-invariant items (%)	Items
Analogies	15 (47%)	4, 7, 8, 9, 10, 14, 16, 17, 21, 22, 24, 27, 28, 30, 31
Arithmetic	2 (13%)	5,9
Computation	9 (41%)	3, 8, 9, 10, 13, 14, 17, 19, 22
Information	19 (48%)	2, 4, 5, 8, 10, 11, 13, 14, 15, 18, 21, 22, 23, 24, 25, 27, 29, 30, 34
Logical Selections	9 (38%)	3, 5, 11, 19, 20, 21, 22, 23, 24
Sentence Completion	7 (35%)	5, 7, 11, 12, 17, 18, 20
Synonyms–Antonyms	13 (33%)	11, 15, 19, 21, 26, 30, 31, 33, 35, 37, 38, 39, 40
Vocabulary	14 (35%)	4, 14, 16, 19, 20, 22, 24, 25, 26, 29, 35, 37, 38, 40

column of Table 3 shows the ES per year (ES/Year), i.e., the FE per year for the 72 years between 1933 and 2006.

While there was a decrease in the variability of θ in all the subtests, many of the decreases were negligible in size. Of the larger decreases, the Information and Synonyms–Antonyms subtests had the largest, dropping from 1.14 in the 1933 sample to 0.65 and 0.63, respectively, in the 2006 sample. The Logical Selections and Vocabulary subtests showed small-to-moderate decreases in variability, dropping .24 and .22 units, respectively.

The reliability of the measures at both time points was very similar. Not only was the mean θSE almost identical for a given subtest across both time points, but the SDs of the θSE were also very similar. Thus, whatever the subtests are measuring, they appear to be measuring it with almost equal precision at both time points.

The traditional (CTT) analysis showed similar effect sizes as the IRT analysis, except for the Information subtest. The IRT analysis showed a moderate FE, while the CTT analysis showed a moderate reverse FE.

4. Discussion

The purpose of the current study was to examine the Flynn effect (FE) in the Estonian version of the National Intelligence Test (NIT; Haggerty et al., 1919) using item response theory (IRT) models. Using data collected from respondents in 1933 and 2006, we asked the following research questions:

1. Do the NIT subtests exhibit at least partial measurement invariance over time in Estonia?
2. If there is at least partial measurement invariance, then do the subtests exhibit a FE?

To answer the first question, the results from the current study show that all the NIT subtests examined have a majority of items that are invariant over time, and that the invariant items were distributed across a wide range of ability levels. Thus, it appears that NIT subtests exhibit partial invariance. To answer the second question, the results showed that seven of the eight individual subtests exhibited a FE and one subtest (Computation) exhibited a reverse FE (see Table 3). For the subtests showing a FE, there was much variability in the magnitude, ranging from an Hedges (1981) g effect size (ES) of 0.24 (3.60 IQ points) on the Arithmetic subtest to an ES of 1.05 (15.75 IQ points) on the Synonyms–Antonyms subtest.

There was not really a pattern of what content domain showed the largest FE, except that the two measures of Gq (i.e., the breadth and depth of a person's acquired store of declarative and procedural quantitative or numerical knowledge; Newton & McGrew, 2010) showed the lowest FE (Arithmetic) and the only reverse FE (Computation). The other subtests, which measure Comprehension–Knowledge and Fluid reasoning, all had both medium and large effect sizes. The CTT results were similar to those from the IRT analysis, except for the Information subtest, where there was a stark contrast in results. The CTT-based scores showed a moderate decrease (−0.33), but the IRT-based scores showed a moderate increase (0.44). There is no reason why the results from the two methods for this particular subtest should differ in their direction, as the content is similar to that from other subtests (e.g., Vocabulary, Synonyms–Antonyms). One possible reason for the disparity between the method results could be the large number of items that were identified as exhibiting DIF for this subtest (19/40 items, approximately 48%). Future studies should examine the impact on number of DIF items and the differences in FE results between the two methods.

In addition to examining mean differences, this study also examined differences in score variability. There was a decrease in the variability of θ in all the subtests, but many of the decreases were negligible in size. There were four exceptions, however. Information and Synonyms–Antonyms subtests had the largest decrease, dropping from 1.14 in the 1933 sample to 0.65 and 0.63, respectively, in the 2006 sample. The Logical Selections and Vocabulary subtests showed small-to-moderate decreases in variability, dropping .24 and .22 units, respectively. One possible explanation for the decrease in variances across the subtests is that the people of Estonia are becoming more homogeneous, meaning there are less distinct groups of people.

Another possible explanation for the decrease in score variability is the difference in sampling methods used in the 1933 and 2006 groups. Must et al. (2003b) have previously criticized Tork's (1940) sample as not being an accurate representation of the population, being biased toward one region of the country. In the current study, there was a slight decrease in age variability from 1933/36 (σ: 1.31 years) to 2006 (σ: .93 years), a difference of 0.38 years or approximately 4.5 months. Other FE studies that have not conditioned on age, however, have shown similar differences. For example, Kanaya and Ceci (2011) had 1–3 month difference in age SD for the time comparison groups. Likewise, Zhou, Zhu, and Weiss (2010) had 1–7 month difference in age SD for the time comparison groups.

One argument against the FE being a “real” change in ability is that most FE studies, to date, have not used appropriate methods that could separate test effects from individual effects. Using IRT methods, this study was specifically able to do that for the NIT subtests. First, this study found there

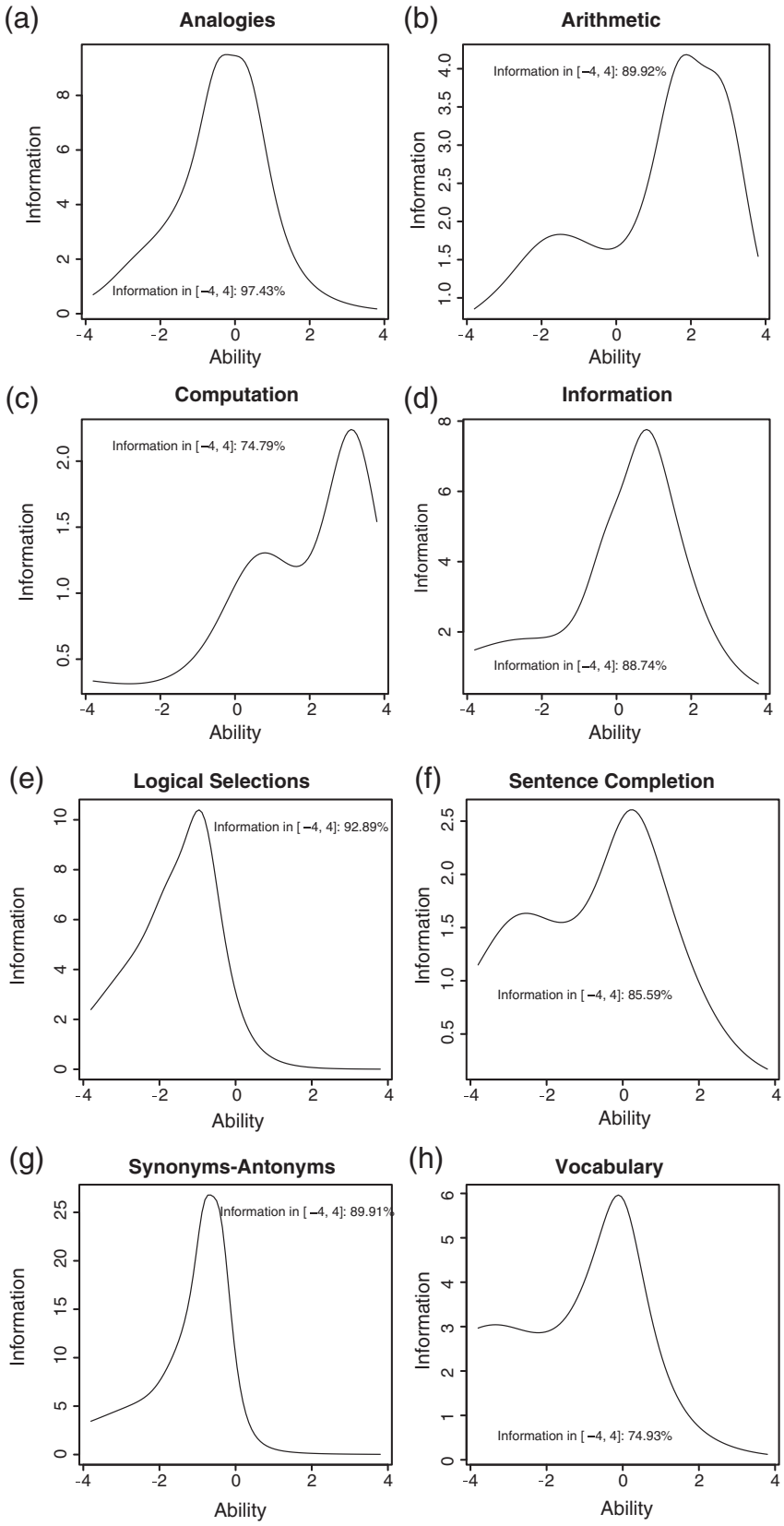

Fig. 2. Test information curves for invariant items.

Table 3
Ability estimate comparisons.

Subtest	Group	*n*	θ		θ Standard Error		Summed Score		IRT		CTT	
			Mean	SD	Mean	SD	Mean	SD	ES	ES/Year	ES	ES/Year
Analogies	1933	890	−0.43	0.87	0.29	0.07	13.54	6.08	1.02	0.014	1.08	0.015
	2006	913	0.45	0.86	0.30	0.09	19.99	5.8				
Arithmetic	1933	890	−0.1	0.88	0.53	0.03	7.72	2.35	0.24	<0.000	0.12	0.002
	2006	913	0.1	0.81	0.52	0.03	7.99	2.24				
Computation	1933	890	0.04	0.84	0.57	0.09	11.73	2.47	−0.10	−0.001	−0.33	−0.005
	2006	913	−0.04	0.77	0.61	0.06	10.93	2.29				
Information	1933	890	−0.21	1.14	0.33	0.07	11.73	2.47	0.44	0.006	−0.33	−0.005
	2006	913	0.20	0.65	0.31	0.05	10.93	2.29				
Logical Selections	1933	890	−0.33	0.95	0.32	0.07	16.42	4.32	0.82	0.001	1.19	0.016
	2006	913	0.36	0.71	0.44	0.15	20.81	2.97				
Sentence Completion	1933	890	−0.27	0.87	0.41	0.08	12.35	3.38	0.64	0.001	0.96	0.013
	2006	913	0.27	0.81	0.49	0.05	15.34	2.88				
Synonyms–Antonyms	1933	890	−0.47	1.14	0.22	0.16	28.84	8.65	1.05	0.002	1.00	0.014
	2006	913	0.49	0.63	0.35	0.14	35.5	3.89				
Vocabulary	1933	890	−0.34	0.97	0.36	0.06	28.54	5.07	0.79	0.011	0.74	0.010
	2006	913	0.34	0.75	0.37	0.07	31.95	4.15				

Note. *n*: sample size, SD: standard deviation; ES: Hedges' (1981) *g* effect size: ES/Year = $\frac{ES}{2006-1933}$ I; RT: item response theory; and CTT: classical test theory.

were some changes in the test properties for all the subtests, with approximately 1/3 to 1/2 of the subtests' items not showing invariance (see Table 2). Consequently, the FE magnitudes found in this study are those after accounting for the changing items. Second, this study examined the reliability of the measurement at both time points. IRT models recognize that reliability of a given measure's score is not the same throughout the distribution of θ, so estimates precision using item and test information (Hambleton & Swaminathan, 1985). Thus, reliability (more specifically, standard error) is estimated separately for ability level, with a lower standard error (θ SE) indicating greater parameter precision and thus higher "reliability." In Table 3 the mean θ SE is very similar for a given subtest across both time points, indicating that, on average, the subtests were measuring θ with equal precision. Moreover, the SD of the θ SE scores is also very similar for a given subtest across both time points, indicating that the subtests were measuring θ with similar precision across the range of θ values. Consequently, the FEs measured in this study do not appear to be due to the test properties changing across time periods.

The results of this study provide additional insight into both the FE and the use of IRT in the role of analyzing the FE. This study was one of the first studies to measure the FE after controlling for changing item properties. While other studies have examined the FE using IRT (e.g., Beaujean & Osterlind, 2008; Beaujean & Sheng, 2010), they only used one or two specific tests, and those tests tended to be specialized measures of some form of achievement (i.e., reading, math). Moreover, this study also was one of the first to specifically examine the precision of the measurements across time points to examine possible changes in the measures across time.

This research is only a start of what needs to be accomplished in this field, though. Future item-level studies of the FE are needed, using both other populations and other measures of cognitive ability. Item level data allows researchers a more precise look at item and test properties, as well as affords them a way to control for items measuring differently across time. While the findings from the current study show that IRT adds a robust method to the analysis of the FE, there are other methods of examining invariance as well as accounting for items that are not invariant.

References

Aguerri, M., Galibert, M., Attorresi, H., & Prieto Maranon, P. (2009). Erroneous detection of nonuniform DIF using the Breslow–Day test in a short test. *Quality and Quantity*, *43*(1), 35–44. http://dx.doi.org/10.1007/s11135-007-9130-2.

Amthauer, R. (1953). *I-S-T. Der Intelligenz-Struktur-test.* GöttingenHogrefe.

Angoff, W. H., & Ford, S. F. (1973). Item–race interaction on a test of scholastic aptitude. *Journal of Educational Measurement*, *10*, 95–105. http://dx.doi.org/10.1111/j.1745-3984.1973.tb00787.x.

Baker, F. B. (1984). Ability metric transformations involved in vertical equating under item response theory. *Applied Psychological Measurement*, *8*(3), 261–271. http://dx.doi.org/10.1177/014662168400800302.

Ballard, P. B. (1922). *Group tests of intelligence.* London, U.K.Hodder and Stoughton.

Bartholomew, D., Knott, M., & Moustaki, I. (2011). *Latent variable models and factor analysis: A unified approach* (3rd ed.)New YorkJohn Wiley Sons.

Beaujean, A. A., & Osterlind, S. J. (2008). Using item response theory to assess the Flynn effect in the national longitudinal study of youth 79 children and young adults data. *Intelligence*, *36*(5), 455–463. http://dx.doi.org/10.1016/j.intell.2007.10.004.

Beaujean, A. A., & Sheng, Y. (2010). Examining the Flynn effect in the general social survey vocabulary test using item response theory. *Personality and Individual Differences*, *48*(3), 294–298. http://dx.doi.org/10.1016/j.paid.2009.10.019.

Beaujean, A. A., & Sheng, Y. (2013). *Assessing the Flynn effect in the Wechsler scales.* (submitted for publication).

Blair, C., Gamsonb, D., Thornec, S., & Bakerd, 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.

Bock, R. D., & Moustaki, I. (2006). Item response theory in a general framework. In C. R. Rao, & S. Sinharay (Eds.), *Handbook of statistics*, *Vol. 26*. (pp. 469–513). Elsevier.

Brand, C. R. (1987). Bryter still and bryter? *Nature*, *328*(6126). http://dx.doi.org/10.1038/328110a0.

Byrne, B. M., Shavelson, R. J., & Muthen, B. O. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, *105*(3), 456–466. http://dx.doi.org/10.1037/0033-2909.105.3.456.

Ceci, S. J., Scullin, M., & Kanaya, T. (2003). The difficulty of basing death penalty eligibility on IQ cutoff scores for mental retardation. *Ethics Behavior*, *13*(1), 11–17. http://dx.doi.org/10.1207/s15327019eb130103.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(2), 233–255. http://dx.doi.org/10.1207/S15328007SEM09025.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.)Mahwah, NJErlbaum.
Cohen, D., Flament, M., Dubos, P. F., & Basquin, M. (1999). Case series: catatonic syndrome in young people. *Journal of the American Academy of Child and Adolescent Psychiatry*, *38*(8), 1040–1046.
Covin, T. M. (1977). Comparison of WISC and WISC-R full scale IQs for a sample of children in special education. *Psychological Reports*, *41*, 237–238.
Dorans, N. J. (2004). Equating, concordance, and expectation. *Applied Psychological Measurement*, *28*(4), 227–246. http://dx.doi.org/10.1177/0146621604265031.
Dorans, N. J., & Kulick, E. (1983). *Assessing unexpected differential item performance of female candidates on SAT and TSWE forms administered in December 1977: An application of the standardization approach (Tech. Rep. No. ETS Research Rep. RR-83-9)*. Princeton, NJEducational Testing Service.
Dorans, N. J., & Kullick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement*, *23*, 355–368. http://dx.doi.org/10.1111/j.1745-3984.1986.tb00255.x.
Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin*, *101*(2), 171–191. http://dx.doi.org/10.1037/0033-2909.101.2.171.
Flynn, J. R. (2006). Tethering the elephant: Capital cases, IQ, and the Flynn effect. *Psychology, Public Policy, and Law*, *12*(2), 170–189. http://dx.doi.org/10.1037/1076-8971.12.2.170.
Flynn, J. R. (2007). *What is intelligence? Beyond the Flynn effect*. New YorkCambridge University.
Flynn, J. R., & Rossi-Casé, L. (2012). IQ gains in Argentina between 1964 and 1998. *Intelligence*, *40*(2), 145–150. http://dx.doi.org/10.1016/j.intell.2012.01.006.
Fraser, C., & McDonald, R. P. (1988). NOHARM: Least squares item factor analysis. *Multivariate Behavioral Research*, *23*, 267–269.
Haertel, E. H. (2006). Reliability. In R. H. Brennan (Ed.), *Educational measurement* (pp. 64–110) (4th ed.). Westport, CTPraeger.
Haggerty, M. E., Terman, L. M., Thorndike, R. L., Whipple, G. M., & Yerkes, R. M. (1919). *National intelligence tests: Manual of directions*. New YorkWorld Book.
Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MAKluwer.
Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*, 107–128. http://dx.doi.org/10.3102/10769986006002107.
Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel–Haenszel procedure. In H. Wainer, & H. I. Braun (Eds.), *Test validity* (pp. 129–148). Hillsdale, NJLawrence Erlbaum.
Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, *18*, 117–144.
Howell, D. C. (2008). Best practices in analysis of variance. In J. W. Osborne (Ed.), *Best practices in quantitative methods*. Los AngelesSage (chap. 23).
Hu, L. -T., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to under parameterized model misspecification. *Psychological Methods*, *3*(4), 424–453. http://dx.doi.org/10.1037/1082-989x.3.4.424.
Jensen, A. R. (1998). *The g factor: The science of mental ability*. Westport, CTPraeger Publishers/Greenwood.
Kanaya, T., & Ceci, S. J. (2007). Mental retardation diagnosis and the Flynn effect: General intelligence, adaptive behavior, and context. *Child Development Perspectives*, *1*(1), 62–63. http://dx.doi.org/10.1111/j.1750-8606.2007.00013.x.
Kanaya, T., & Ceci, S. J. (2011). The Flynn effect in the WISC subtests among school children tested for special education services. *Journal of Psychoeducational Assessment*, *29*(2), 125–136. http://dx.doi.org/10.1177/0734282910370139.
Kanaya, T., Scullin, M. H., & Ceci, S. J. (2003). The Flynn effect and U.S. policies: The impact of rising IQ scores on American society via mental retardation diagnoses. *American Psychologist*, *58*, 778–790. http://dx.doi.org/10.1037/0003-066X.58.10.778.
Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement, Vol. 4th*. (pp. 17–64)Westport, CTPraeger.
Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.)New YorkSpringer-Verlag.
Little, T. D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioral Research*, *32*(1), 53–76. http://dx.doi.org/10.1207/s15327906mbr32013.
Lynn, R. (2009). Fluid intelligence but not vocabulary has increased in Britain, 1979–2008. *Intelligence*, *37*(3), 249–255. http://dx.doi.org/10.1016/j.intell.2008.09.007.
Lynn, R., Allik, J., Pullman, H., & Laidra, K. (2002). A study of intelligence in Estonia. *Psychological Reports*, *91*(3), 1022–1026. http://dx.doi.org/10.2466/pr0.2002.91.3.1022.
Lynn, R., Pullmann, H., & Allik, L. (2003). A new estimate of the IQ in Estonia. *Perceptual and Motor Skills*, *97*(2), 662–664. http://dx.doi.org/10.2466/pms.2003.97.2.662.
Magis, D., Beland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods*, *42*(3), 847–862. http://dx.doi.org/10.3758/BRM.42.3.847.
McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJErlbaum.
McGrew, K. S. (2010). The Flynn effect and its critics: Rusty linchpins and "Lookin' for g and Gf in some of the wrong places". *Journal of Psychoeducational Assessment*, *28*(5), 448–468. http://dx.doi.org/10.1177/0734282910373347.
Meredith, W., & Teresi, J. A. (2006). An essay on measurement and factorial invariance. *Medical Care*, *44*(11), S69–S77. http://dx.doi.org/10.1097/01.mlr.0000245438.73837.89.
Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (pp. 13–103) (3rd ed.). Washington, D. C.American Council on Education.
Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New YorkRoutledge.
Mingroni, M. A. (2007). Resolving the IQ paradox: Heterosis as a cause of the Flynn effect and other trends. *Psychological Review*, *114*, 806–829.
Must, O., & Allik, J. (2011). Intelligentsuse uurimine Eestis. In R. Mõttus, J. Allik, & A. Realo (Eds.), *Intelligentsuse psühholoogia* (pp. 344–364). Tartu, EstoniaTartu Ülikooli Kirjastus.
Must, O., & Must, A. (2013). Changes in test-taking patterns over time. *Intelligence*, *41*(6), 780–790 (this issue).
Must, O., Must, A., & Raudik, V. (2003a). The Flynn effect for gains in literacy found in Estonia is not a Jensen effect. *Personality and Individual Differences*, *34*(7), 1287–1292. http://dx.doi.org/10.1016/s0191-8869(02)00115-0.
Must, O., Must, A., & Raudik, V. (2003b). The secular rise in IQs: In Estonia, the Flynn effect is not a Jensen effect. *Intelligence*, *31*(5), 461. http://dx.doi.org/10.1016/S0160-2896(03)00013-8.
Must, O., te Nijenhuis, J., Must, A., & van Vianen, A. E. M. (2009). Comparability of IQ scores over time. *Intelligence*, *37*(1), 25–33. http://dx.doi.org/10.1016/j.intell.2008.05.002.
Muthen, L. K., & Muthen, B. O. (2010). *Mplus (version 6) [computer software]*. [computer software]. Los Angeles, CAMuthén and Muthén.
National Research Council (1920). *National intelligence tests…revised for English schools*. London, U.K.Harrap.
Neisser, U. (Ed.). (1998). *The rising curve: Long-term gains in IQ and related measures*. Washington, DCAmerican Psychological Association.
Newton, J. H., & McGrew, K. S. (2010). Introduction to the special issue: Current research in Cattell–Horn–Carroll-based assessment. *Psychology in the Schools*, *47*(7), 621–634. http://dx.doi.org/10.1002/pits.20495.
Penfield, R. (2003). Applying the Breslow–Day test of trend in odds ratio heterogeneity to the analysis of nonuniform DIF. *Alberta Journal of Educational Research*, *49*(3), 232–243.
Pietschnig, J., Voracek, M., & Formann, A. K. (2010). Pervasiveness of the IQ rise: A cross-temporal meta-analysis. *PLoS One*, *5*(12), e14406. http://dx.doi.org/10.1371/journal.pone.0014406.
Ployhart, R. E., & Vandenberg, R. J. (2010). Longitudinal research: The theory, design, and analysis of change. *Journal of Management*, *36*(1), 94–120. http://dx.doi.org/10.1177/0149206309352110.
Pullmann, H., Allik, J., & Lynn, R. (2004). The growth of IQ among Estonian schoolchildren from ages 7 to 19. *Journal of Biosocial Science*, *36*(6), 735–740. http://dx.doi.org/10.1017/s0021932003006503.
R Development Core Team (2012). *R: A language and environment for statistical computing*. Vienna, AustriaR Foundation for Statistical Computing (Available from http://www.R-project.org).
Raven, J. C. (1958). *Standard progressive matrices*. Cambridge, UKThe University Press.
Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software*, *17*(5), 1–25.
Rodgers, J. L. (1998). A critique of the Flynn Effect: Massive IQ gains, methodological artifacts, or both? *Intelligence*, *26*(4), 337–356. http://dx.doi.org/10.1016/s0160-2896(99)00004-5.
Sanborn, K. J., Truscott, S. D., Phelps, L., & McDougal, J. L. (2003). Does the Flynn effect differ by IQ level in samples of students classified as learning disabled? *Journal of Psycheducational Assessment*, *21*, 145–159 (Journal Article).
Sigman, M. D., & Whaley, S. E. (1998). The role of nutrition in the development of intelligence. In U. Neisser (Ed.), *The rising curve: Long-term gains in IQ and related measures* (pp. 155–182). Washington, D.C.American Psychological Association.
Steenkamp, J. -B. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, *25*(1), 78–107.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*, 361–370. http://dx.doi.org/10.1111/j.1745-3984.1990.tb00754.x.

te Nijenhuis, J., Cho, S. H., Murphy, R., & Lee, K. H. (2012). The Flynn effect in Korea: Large gains. *Personality and Individual Differences*, *53*(2), 147–151. http://dx.doi.org/10.1016/j.paid.2011.03.022.

te Nijenhuis, J., Murphy, R., & van Eeden, R. (2011). The Flynn effect in South Africa. *Intelligence*, *39*(6), 456–467. http://dx.doi.org/10.1016/j.intell.2011.08.003.

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

Thompson, M. S., & Green, S. B. (2006). Evaluating between-group differences in latent variable means. In G. R. Hancock, & R. O. Meuller (Eds.), *Structural equation modeling: A second course* (pp. 119–169). Greenwich, CTInformation Age.

Toim, K. (1976). The use of Raven's progressive matrices test for measurement of mental developments of pupils. *Acta et Commentationes Universitatis Tartuensis*, *395*, 53–59.

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

van der Linden, W. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*(3), 287–308. http://dx.doi.org/10.1007/s11336-006-1478-z.

Vandenberg, R. J. (2002). Toward a further understanding of and improvement in measurement invariance methods and procedures. *Organizational Research Methods*, *5*(2), 139–158. http://dx.doi.org/10.1177/1094428102005002001.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*(1), 4–70. http://dx.doi.org/10.1177/109442810031002.

Wai, J., & Putallaz, M. (2012). The Flynn effect puzzle: A 30-year examination from the right tail of the ability distribution provides some missing pieces. *Intelligence*, *39*(6), 443–455. http://dx.doi.org/10.1016/j.intell.2011.07.006.

Whipple, G. M. (1921). The national intelligence tests. *The Journal of Educational Research*, *4*(1), 16–31.

Yerkes, R. M. (1921). *Memoirs of the national academy of sciences: Psychological examining in the United States Army*, *Vol. 15*, Washington, D. C.Government Printing Office.

Yoo, B. (2002). Cross-group comparisons: A cautionary note. *Psychology and Marketing*, *19*(4), 357–368. http://dx.doi.org/10.1002/mar.10014.

Young, B., Boccaccini, M. T., Conroy, M. A., & Lawson, K. (2007). Four practical and conceptual assessment issues that evaluators should address in capital case mental retardation evaluations. *Professional Psychology: Research and Practice*, *38*(2), 169–178.

Zhou, X., Zhu, J., & Weiss, L. G. (2010). Peeking inside the “Black Box” of the Flynn Effect: Evidence from three Wechsler instruments. *Journal of Psychoeducational Assessment*, *28*(5), 399–411. http://dx.doi.org/10.1177/0734282910373340.