Intelligence 41 (2013) 780–790

Contents lists available at ScienceDirect

Intelligence

Changes in test-taking patterns over time

Olev Must [a],*, Aasa Must [b]

[a] University of Tartu, Estonia
[b] Estonian National Defense College, Estonia

ARTICLE INFO

Article history:
Received 13 February 2013
Received in revised form 12 April 2013
Accepted 13 April 2013
Available online 15 May 2013

Keywords:
Flynn Effect
National Intelligence Tests
Test-taking pattern
Tork
Estonia

ABSTRACT

The current study aims to investigate the relationship between right, wrong and missing answers to cognitive test items (test-taking patterns) in the context of the Flynn Effect (FE). We compare two cohorts of Estonian students (1933/36, $n = 890$; 2006, $n = 913$) using an Estonian adaptation of the National Intelligence Tests and document three simultaneous trends: fewer missing answers (−1 Cohen's *d* averaged over subtests), and a rise in the number of right and wrong answers to the subtests (average *d*s of .86 and .30, respectively). In the Arithmetical Reasoning and Vocabulary subtests, adjustments for false-positive answers (the number of right minus the number of wrong answers) reduced the size of the Flynn Effect by half. These subtests were supposed to be high *g*-loading subtests. Our conclusion is that rapid guessing has risen over time and influenced tests scores more strongly over the years. The FE is partly explained by changes in test-taking behavior over time.

1. Introduction

Both cognitive and non-cognitive factors may affect intelligence (IQ) test performance. The perceived consequences of the test and the test-takers' desire to do their best to pass the test or score highly may influence the final result. These factors are probably not constant over time and may be related to changes in the test score over time. The current study focuses on the changes in test-taking patterns — the relationships between right, wrong and missing answers and their contribution to changes in overall test score over time.

The Flynn Effect (FE) is the rise of IQ test scores between different historical cohorts (Flynn, 2012). Several explanations for the effect have been offered. These can be largely divided into two parts — does the rise in IQ test scores indicate an enhancement of latent cognitive abilities over the decades, or is it connected to some other developments which render the effect hollow.

The term Jensen Effect (Rushton, 1999) was coined to denote the importance of correlates to the rise in IQ test scores. The Jensen Effect suggests that if we are dealing with a real improvement in the general factor (*g*) of intelligence, then changes in mean scores on subtests should correlate with the degree to which each subtest correlates with the *g* factor, i.e., with their *g* loading (Jensen, 1998; te Nijenhuis & van der Flier, 2013). *g*-Differences have been found to be the best predictors of a variety of variables, including not only scholastic and workplace performances, but also of brain size, brain pH, brain glucose metabolic rate, average evoked potential, reaction time, and other physiological factors (see Jensen, 1980, 1998). The degree to which the *g* factor correlates with gains in subtest scores is employed as a yardstick for deciding the meaning of the rise in IQ test scores. Rushton (1999) was the first to present evidence that the secular rise in IQ was not connected with *g*. Must, Must, and Raudik (2003) also found that the gains in subtest scores were not correlated with the subtests' *g* loadings and that the Flynn Effect is not the Jensen Effect. At the same time there have been alternative findings that positive correlations between *g* and FE also exist (e.g. Colom, Juan-Espinosa, & Garcia, 2001).

Investigators of the measurement properties of IQ tests argue that the comparisons of test score averages over time are meaningless if the measures are not invariant. The same tests may have different measurement characteristics over a

* Corresponding author.
E-mail address: Olev.Must@ut.ee (O. Must).

0160-2896/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.intell.2013.04.005

long time period. Wicherts et al. (2004) showed that in every dataset they studied, measurement invariance was absent, i.e. students of the same ability level in different cohorts achieved different observed test scores.

An absence of measurement invariance can be attributed to a number of factors, including test anxiety (e.g., stereotype threat, Wicherts, Dolan, & Hessen, 2005), item meaning (Must, te Nijenhuis, Must, & van Vianenen, 2009), or test-specific abilities (see Wicherts & Dolan, 2010).

The main aim of the investigation into *g*-correlates of FE and testing measurement invariance is to estimate the significance and reality of the phenomenon. There are some additional constructs that may influence the testing process, and these can also change over time. Primarily we refer to the influence of the emotional states of the test-takers and their attitudes towards the results of testing and their approach to taking the test.

In the initial stages of IQ testing development, it was assumed that speed and accuracy need not be distinguished within the IQ domain. In the last decades, however, this has changed and these facets have garnered more scrutiny. Speed and power have become two of the fundamental measures to be considered when analyzing cognitive tests. The constructs are ambiguous because they can easily vary for the same task and for the same person. It should be noted that a trade-off strategy of sacrificing accuracy for the sake of speed may also have an additional impact (e.g., Furneaux, 1960; Klein Entink, Kuhn, Hornke, & Fox, 2009; Partchev, De Boeck, & Steyer, 2011; Rindler, 1979; van der Linden, 2011; van der Maas, Molneaar, Maris, Kievit, & Borsboom, 2011).

Test anxiety was probably the first non-cognitive covariate to cognitive measures that was researched thoroughly (Pintrich, Cross, Kozma, & McKeachie, 1986). Recently Wicherts and Zand Scholten (2010) showed that the test anxiety of test-takers may affect the validity of cognitive tests. The influence of test-taking motivation on educational test performance is clearly documented (Barry, 2010; Baumert & Demmrich, 2001; Eklöf, 2010; Sundre & Kitsantas, 2004; Wise & DeMars, 2005, 2010). Duckworth, Quinn, Lynam, Loeber, and Stouthamer-Loeber (2011) investigated the role of test-taking motivation in intelligence testing and concluded that under low-stakes conditions (when test results have no consequences), some individuals try harder than others. Test motivation can act as a confounding factor that inflates estimates of the predictive validity of intelligence in relation to life outcomes (Wicherts & Zand Scholten, 2010). In FE research, the works of Wise and Kong (2005) and Wise, Pastor, and Kong (2009) regarding the response time effort (RTE) in test-taking and the development of the concept of rapid-guessing behavior in low-stakes testing are especially valuable. The construct of RTE is based on the hypothesis that when administered an item, low motivated examinees will answer quickly without allowing themselves enough time to read and fully consider the item.

RTE is related to computer-based testing procedures. Computers make it possible to measure RTE along with conventional cognitive tests. The RTE measure is analogous to traditional reaction or inspection time measures (Bates & Eysenck, 1993; Deary & Stough, 1996; Lee & Chen, 2011), as the focus of attention is on the speed of reaction and not so much on its quality. Bates and Eysenck (1993) argue that inspection time measurements do not make the motivational and response strategy factors of the measurement process visible. Rapid guessing in test-taking means that test-takers respond without actually giving their best effort, thus the decision time is too short and effort too minimal to solve the problems. The result of the rapid guessing strategy is that the test-takers do not perfectly adhere to the testing requirements and too often respond randomly, and thus make many errors. However, by chance rapid guessing also generates a greater number of correct answers, depending on the number of answer options in a multiple choice problem. Similar to RTE is the concept of mental taxation. Wolf, Smith, and Birnbaum (1995) demonstrated that the solving of mentally taxing items is highly informative for categorizing test-takers based on test-taking motivation. Wolf et al. described mental taxation in terms of the experts' evaluations of the work necessary to solve a certain item rather than the amount of time the test-takers used for the item. The theoretical background of mental taxation for items is taken from the expectancy–value theory of motivation (Atkinson, 1957; Eccles & Wigfield, 2002). The expectancy component can be divided into two parts: the first is concerned with the likelihood of success, and the second is concerned with the effort necessary to arrive at a correct answer.

Brand (1996) applied the idea of rapid-guessing to the comparison of test results from different time periods. He noticed that FE "….evidence is drawn largely from short, timed, multiple-choice, group-administered tests of IQ on which there is no adjustment for guessing. Scores on such tests may have improved since 1945 not just because of rising *g* levels but because of modern educators' encouragements to children to avoid 'obsessional' accuracy and 'pedantic' attention to detail. Being composed of different sections, each requiring the use of different principles (e.g. series completion, analogies, oddity), most group tests effectively penalize test-takers who strive for accuracy. Such testees spend valuable time trying to be quite sure they are giving correct answers — rather than making use of guesswork" (Brand, 1996, p. 140). The same inference regarding changes in test-taking patterns over time was proposed by Wicherts et al. (2004) after analyzing several FE data sets.

The historical data used for the estimation of FE, including considerations of test-taking behavior and the fixing of corresponding variables, was not originally collected with the intention of utilizing these data decades later. Fortunately, however these data include some elements that allowed for the estimation of the influence of factors related to the test-taking process. Historical data fixed at the item level allow us to investigate the changes in the relationships between right, wrong and missing responses and the influence of the item presentation order on the subtest.

Various sources have shown substantial differences between power and speed dimensions in IQ test performance (eg. Partchev et al., 2011; van der Linden, 2011); these findings concerning the differences in mental taxation of items (Wolf et al., 1995), response time effort and rapid guessing (Wise & DeMars, 2005; Wise & Kong, 2005; Wise et al., 2009), as well as perceived consequences of the test results on the test-takers (Wolf & Smith, 1995) and concepts regarding changes in test-taking patterns over time (Brand, 1996, Wicherts et al., 2004) allow us to propose the following hypothesis for the current investigation:

The FE – the rise in IQ test scores over time – is partly explained by the changes in test-taking behavior.

This general assumption can be divided into the following problems/sections.

1. If FE means a real improvement in latent cognitive ability of later cohorts, then there should be a rise in only right answers to the items (in the case of all other conditions being invariant). The rise of wrong answers over time is a clear sign of a negative FE (Lynn & Harvey, 2008) or, if the number of both right and wrong answers is simultaneously on the rise, then it is a sign of a rise in guessing behavior. The rise of both right and wrong answers is possibly due to there being fewer missing answers. The first research question is then — what are the secular trends of different answers over time?
2. The relationship between wrong and right answers may vary across subtests. This is to be expected as the subtests can (1) tap different latent cognitive abilities, (2) involve a different number of items, (3) be administered under different time limits, (4) have various answer options, and (5) have different item characteristics (more or less difficult and discriminating).
3. The simultaneous rise of right and wrong answers implies that some of the right answers may be interpreted as false-positive — they are the result of guessing. How much can the adjustment of test results (right minus wrong answers) influence the rise in test scores?
4. The item presentation order in subtests may be related to the changes in test-taking response patterns. If over the years test-taker's strategies have shifted towards providing more quick responses, then this should be more visible at the end of a subtest — in the case of speeded tests, test-takers may simply hurry more to finish the test. It can be posited that over time the answers that were arrived at by chance as well as wrong answers are correlated with the item presentation order.

2. Method

2.1. National Intelligence Tests

The National Intelligence Tests (NIT) (Haggerty, Terman, Thorndike, Whipple, & Yerkes, 1920; Terman, 1921; Whipple, 1921) were developed in 1919–1920 for the measurement of the intelligence of schoolchildren (grades 3 to 8) by a team of psychologists – M. Haggerty, L. Terman, E. Thorndike , G. Whipple and R. Yerkes – who had previous experience with American Army Mental Tests (Brigham, 1923; Yoakum & Yerkes, 1920). The team decided to arrange the NIT subtests into two groups (Scales A and B) in two forms. The Scale A consisted of 5 subtests: Arithmetical Reasoning (A1), Sentence Completion (A2), Logical Selection (A3), Same–Opposite (A4) and Symbol–Digit (A5). Scale B involved the following subtests: Computation (B1), Information (B2), Vocabulary (B3), Analogies (B4) and Comparisons (B5). The NIT resembled the army mental tests (alpha and beta) in several respects. The majority of the NIT subtests were analogous to army subtests, including some of the subtests from the beta version (A5 and B5). The army tests were also administered in an authoritative, militaristic style, and the same was intended for the NIT. “Directions, and especially commands, should be spoken authoritatively, and instant obedience should be expected and required. Every child should obey promptly and without question” (Haggerty et al., 1920, p. 6). Exact timing requirements were especially stressed — there were in total 20 different timing sections in the NIT testing method. “Ready — Go!”, “Stop!”, and “Pencil up!” are frequent commands in the NIT testing manual. An important similarity between the NIT and army tests is the usage of penalties for wrong answers. At the same time penalties are not made explicit for test-takers, but they are used in the test scoring procedure. The manual of the army test explains: “In scoring the tests certain mathematical precautions are necessary”. For instance, the test which offers only two alternatives will yield a high percentage of right scores by chance. To compensate for this, such a test may be scored “right minus wrong” (Yoakum & Yerkes, 1920, p. 6–7). There is one important difference between the administration of the NIT and army tests. The army tests were used in concordance with military discipline with the hope that the test-takers would show their best abilities, and with the awareness that the results would be forwarded to a commanding officer and could be a factor in their careers. The motivation of schoolchildren to take the test was rather general without reference to the personal consequences: “This is a test to find out what things boys and girls can do. You must do your very best, because we want to see whether the children of the….. school can do as well as others. I think you find the test very interesting.”(Haggerty et al., 1920, p. 8).

2.2. Measures of the current study

An Estonian adaptation of the NIT (Tork, 1940) was used. The test was adapted at the beginning of the1930s. Tork utilized the NIT manual but used British versions of the test (Haggerty, Terman, Thorndike, Whipple, & Yerkes, n.d.-a,b). The test design, structure and main body of the nonverbal items were not changed in the adaptation. All verbal subtests were analogs to the original NIT versions. The original test-booklets with answers are stored in the Estonian National Historical Archive. Tork's adaptation (with only minimal changes) of the test was used later (in 1997 and 2006) by Must et al. (2003, 2009) to estimate FE in Estonia. For instance, in the current tests, some of the facts in the Information subtest were updated, such as The great war (World War I) was changed to World War II etc.

2.3. Description of the subtests

A1 Arithmetical Reasoning. The subtest consists of 16 items requiring a solution of one unknown quantity. For example: “How many pence are six pence and five pence?”

A2 Sentence Completion. The subtest consists of 20 items requiring the filling in of a missing word to make sentence understandable and correct. An example: “Time ……sometimes worth more ……money”.

A3 Concepts. The historical name of the subtest (Logical Selection) is not exact, as the items are about the comprehension of various concepts. The subtest consists of 24 items requiring the selection of two characteristic features from among those given. For example “Shoe: button, foot, sole, toe, tongue. Cat: sphere, claws, eating, eyes, mouse”.

A4 Same–Different. The historical subtest name (Same–Opposite) is not correct; the item content deals with the sameness and distinctiveness of concepts. The subtest consists of 40 items requiring the evaluation of whether the meaning of the words presented is the same or different. For example: light….bright, liquid…solid.

A5 Symbol–Digit. The subtest consists of 120 items requiring a decision regarding which digit should be assigned to a symbol based on a key; 9 different symbols were presented. An abridged example of the key:

+	Δ	∞
1	2	3

B1 Computation. The subtest consists of 22 items requiring addition, subtraction, multiplication, and division of both integers and fractions. For example: Subtract: 39260 − 16678.

B2 Information. The subtest consists of 40 items of everyday knowledge. For example: "The days and nights are nearest equal in January/June/March/May".

B3 Vocabulary. The test consists of 40 items requiring knowledge of the qualities of different objects, for example: "Have frogs wings?"

B4 Analogies. The test consists of 32 items requiring the transference of the relation of two given words to other presented words. For example: wolf–sheep–cat–fur, kitten, dog, and mouse.

B5 Comparisons. The test consists of 50 items requiring judgment about the sameness of sets of numbers, family names, and graphic symbols presented in two columns. For example: 5628913653…..5628813653.

Two of the subtests (A1 and B1) required constructed responses to the items; the other subtests employed multiple-choice answer systems. With the exception of subtest A3 (Concepts) all subtests used a 1 point coding system (a correct answer was not divisible into partial credit components). In subtest A3 partial credit means that one correct answer gave 1 point and two correct answers, 2 points. The original scoring system used weights for subtests' scores and penalties for wrong answers. The number of right answers was multiplied by 2 in subtests A1, A2, and B1. The multiplier 3/10 was applied for the subtest B5 (Symbol–Number Comparisons). All scales were speeded, the testing time for a subtest varied from 2 min (Comparison) to 4 min (Computation). Various numbers of items and varying time limits meant that the time per item in subtests varied more than 12 times (1.5 s in subtest A5 up to 18.8 s in subtest A1). The test manual (Tork, 1940) fixed a strict time control for each item type practice. In actual subtests, there were a total of 20 different timing sections. As a rule, the fixed time was not sufficient to allow most test-takers to solve all the items and this caused a considerable number of missing answers. The NIT requires attention, concentration and motivation from the test-takers, who had to follow the examiner's instructions in order to complete the test. Although Tork (1940) did not explain the reasons for the time differences, weights and penalties, all of the technical details of testing are important variables for the current analysis (Table 1). For test-takers the testing instruction did not include information about weights and penalties for wrong answers. At the same time the tests' title page included the background information (gender, age, parents' occupation, school etc.) and the scoring table for test results at the bottom of the page. This table was used by the researchers to make calculations after the testing and it also became a principal source of information about the test-taking performance. Similarly to the original NIT, the Estonian adaptation of 1934 included a table with a scoring algorithm. The scoring algorithm was not included in later tables (in 1997 and 2006). The reason for this change was that in 2006 data preparation and analysis became computerized and manual scoring was not needed. Each subtest started with a brief introduction, which typically did not include motivational statements. The test-takers were instructed to mark/underline the right answer. However, the instruction of the subtest B3 (Vocabulary) suggested that the test-takers – do as many as you can – and the instruction of the subtest B1 (Computation) instructed: "Do this work in arithmetic as quickly as you can without making mistakes. Try each example as you come to it. Look carefully at each one to see what you are to do". This is the only part of the test instruction where references to mistakes or wrong answers are made. Archival data show that the manual scoring was checked twice, but there is no information that the test-takers received later feedback about their test performance.

2.4. Samples and data

The estimation of FE typically means that the comparison of average test results by cohorts is comparable with each other chronologically. This paper utilizes the same samples that were used by Must et al. (2009). The older sample (1933/36; N = 890) consists of students from grades 4 to 6, mean age 13.3 (SD = 1.24) years, and the younger sample (2006, N = 913) from grades 6 to 8 with mean age 13.5 (SD = .93) years. The data for the sample from 1933/36 was taken from the Estonian National Historical Archive (foundation EAA.2101), the sample data for 2006 came from the same region as the historical sample and was collected under the supervision of the authors. Test-takers were of the same age, but there was a difference in schooling of 2 years. The reason for this difference is due to the lowering of the age of obligatory school attendance.

2.5. The logic of analysis

The aim of the current analysis is not the estimation of the size of FE, but rather an estimation of the possible impact of the differences in response patterns of students on their test results.

Estonian historical data allow us to analyze the right, wrong and missing answers, as actual test-books with responses are available. First we present data regarding the changes of corresponding answers over time at the level of means of subtests. We also adjusted the students' test score (sum of right answers), assigning penalties for wrong answers. The term "adjusted score" means that the number of wrong answers is subtracted from the number of right answers. For an estimation of the effect size between the cohorts' scores, the Cohen's *d* is used.

The change in test-taking behavior is also estimated via an odds ratio at the item level.

Odd means the chance of occurrence of the event (odd of the event x = p/(1 − p), where p stands for the probability

Table 1
Time limits and the NIT scoring system.

Subtest	No of items in subtest	Minutes per subtest	Sec per item	Scoring algorithm
Arithmetical Reasoning (A1)	16	5	18.8	Number of right answers × 2
Sentence Completion (A2)	20	4	12	Number of right of answers × 2
Concepts (A3)	24	3	7.5	Partial credit system (1 and 2 points for right answers); summary of credit points
Same–Different (A4)	40	2	3	Number of right answers − number of wrong answers
Symbol–Digit (A5)	120	3	1.5	Number of right answers × 3/10
Computation (B1)	22	4	10.9	Number of right answers × 2
Information (B2)	40	4	6	Number of right answers
Vocabulary (B3)	40	3	4.5	Number of right answers − number of wrong answers
Analogies (B4)	32	3	5.6	Number of right answers
Comparisons (B5)	50	2	2.4	Number of right answers − number of wrong answers

of the event x). The odds ratio (OR) describes changes in the occurrence of an interesting variable in two situations (odd 1/odd 2). The main reason to use OR in the current analysis is that it allows for the investigation of test-taking at the item level. OR at the item level allows one to evaluate the universality of change across used items.

Analysis at the item level renders some advantages for the investigation of changes in response patterns — to estimate the influence of the item presentation order on test-taking behavior. The subtest A3 is not included in this item-level analysis, as this particular, partial credit scoring was used (wrong and missing answers had a different value than in the other subtests).

3. Results

In the period 1933/36–2006 mean subtest results of comparable age-cohorts have changed (Table 2). There is a general pattern that the frequency of missing answers in NIT subtests is diminished (approximately 1 *d*), with the exception of the subtest B1 (Computation), where the rise in missing answers was 0.36 *d*. The rise of right answers is evident in most of the subtests (7 from 9). The mean rise of right answers per subtest is about .86 *d*. The frequency of wrong answers rose as well. The mean rise effect of wrong answers (.30 *d*) is smaller than the mean rise in right answers, but it is also evident in 7 of the 9 subtests. In the FE framework it is important to note that the diminishing number of missing answers is offset by, not only right answers, but wrong answers as well.

Over time the general relationship between right, wrong and missing answers has changed.

One of the clearest findings in both cohorts is that instead of right answer there are missing answers. This correlation between the number of correct answers and missing answers was more apparent in 1933/36 ($r = -.959$, $p < .001$) than in 2006 ($r = -.872$, $p < .001$). In 1933/36 the number of wrong answers did not correlate with the number of right answers ($r = -.005$), but in 2006 the frequency of wrong answers moderately indicates a low number of right answers ($r = -.367$, $p < .001$). In both cohorts the number of missing answers is negatively correlated with wrong answers, but the relationship is stronger in the 2006 cohort ($r = -.277$, $p < .001$; in 1933/36 $r = -.086$, $p = .01$). The cohort differences between the above presented correlations across cohorts are statistically significant.

The adjustment for wrong answers has a different effect in different subtests (Fig. 1). The adjustment for mistakes reduced the effect size for the gain on subtest A1 (Arithmetical Reasoning) from $d = .55$ to $d = .15$ and reduced the effect size for the gain on subtest B3 (Vocabulary) from $d = .74$ to $d = .30$. The adjustment minimally reduced the effect size for subtests A5 (Symbol–Digit; from 1.65 to 1.62 *d*) and B5 (Comparisons; from 1.71 to 1.61 *d*). The adjustment did

Table 2
Secular changes in NIT: Estonian student cohorts 1933/36–2006.

	1933/36						2006						Effect size (Cohen's *d*)		
	Right answers		Wrong answers		Missing answers		Right answers		Wrong answers		Missing answers		Right answers	Wrong answers	Missing answers
	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD			
Arithmetical Reasoning (A1)	6.8	2.2	2.5	1.8	5.6	2.4	8	2.2	3.2	2	4.8	2.5	0.55	0.37	−0.33
Sentence Completion (A2)	11.4	3.4	2.3	1.7	5.4	3.1	15.3	2.9	2.9	1.9	1.7	2.1	1.24	0.33	−1.42
Same–Different (A4)	28.8	8.6	2.9	2.6	8.3	8.2	35.5	3.9	3.4	2.6	1.1	2.9	1.07	0.19	−1.30
Symbol–Digit (A5)	74.2	20.6	0.71	4.4	45.1	20.5	105	16.4	0.51	1	14.7	16.4	1.65	−0.07	−1.65
Computation (B1)	11.7	2.5	3	1.8	7.2	2.3	10.9	2.3	3	1.8	8	2.2	−0.33	0.00	0.36
Information (B2)	22.2	7.4	5.8	4.3	11.9	8.7	22.1	4.9	10.8	4.8	7.1	6.5	−0.02	1.10	−0.63
Vocabulary (B3)	28.5	5.1	3.8	2.4	7.7	5.5	31.9	4.1	5.5	2.7	2.6	4	0.74	0.67	−1.07
Analogies (B4)	13.5	6.1	8	5	10.4	6.4	20	5.8	8.1	6	3.9	3.8	1.09	0.02	−1.27
Comparisons (B5)	27.9	6.4	2.1	2.6	19.92	7.2	38.5	6	2.4	2.1	9.1	6.5	1.71	0.13	−1.58
M													0.86	0.30	−0.99

Note. M = mean SD = standard deviation.

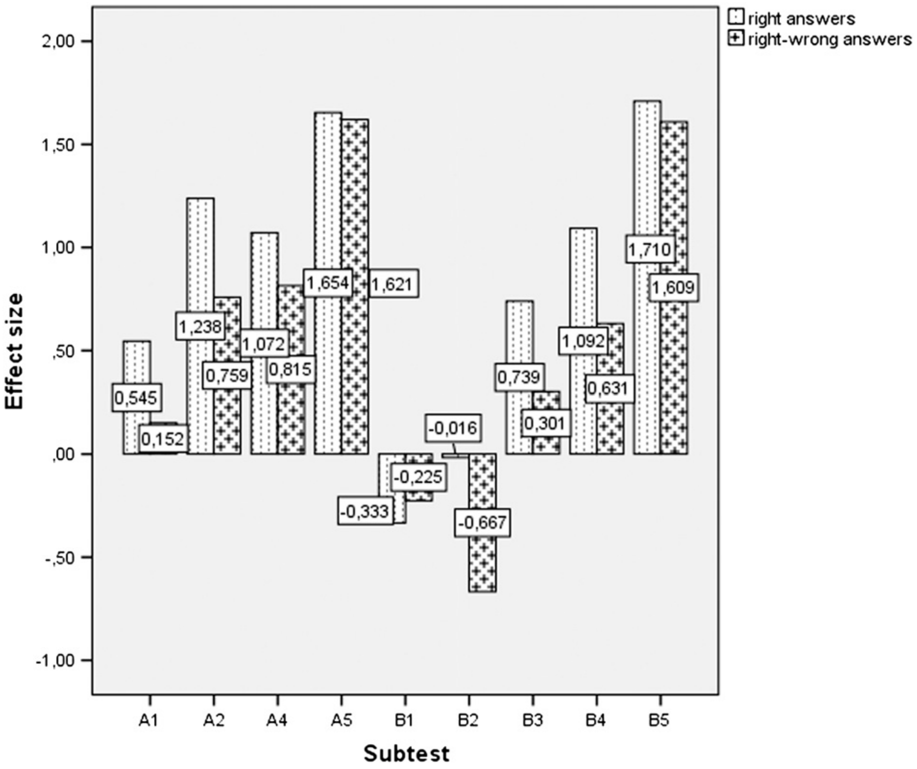

Fig. 1. Secular changes in NIT subtests' mean scores (right and corrected right answers). Cohen's *d*.

not change the decreases of scores in two subtests B1 (Computation) and B2 (Information). Moreover, in the last subtest the negative trend became more apparent (effect sizes correspondently −.02 *d* to −.67 *d*).

The changes in the response format are described not only as differences in subtest averages, but also as changes of odds to give a right, wrong or missing answer at the item level (Table 4). The items may have a different meaning over time, and this can be the reason why the items have a different response structure. These differences in the structure of the items' meaning are reflected by high values of odds ratio (OR) of different responses. It is for this reason that for a conservative evaluation of test answering patterns, the median of OR is a better indicator than the average value. Within the context of the current paper it is important to note that in 7 subtests out of 9, the median OR of wrong answers is over 1. In one subtest (B1) the OR of wrong answers is below 1, and was at the same level as the OR of right answers. It can be concluded that at the item level the odds of both answers – right and wrong – are rising.

Finally, we investigate how wrong answers are dependent on the order of items in the subtest (Table 5). Subtest A3 (Concepts, partial credit system) and subtest B5 (Symbol–Digit; the 120 items of the subtest about symbol–digit correspondence are not entered at the item level) are not included in this analysis. The difference between the two cohorts lies in the dependence of item presentation order on the different response types. In 1933/36 there was no clear pattern of relation between the item presentation order and the odds of a wrong answer at the subtest level. In one subtest (B3, Vocabulary) the correlation was statistically significant and strongly positive ($r = .705$, $p < .01$), and in two others (B4 and B5) it was moderately negative ($r = -.438$, $p < .05$ and $r = -.304$, $p < .05$ accordingly). The 2006 relationship between the odds of wrong answers is correlated positively with item presentation order (correlations varied from $r = .267$,

Table 3
The relationships between right, wrong and missing answers: 1933/36–2006 (Pearson correlation coefficients).

Variable	Number of correct answers	Number of wrong answers	Number of missing answers
Number of correct answers		−.005	−.959**
Number of wrong answers	−.367**		−.277**
Number of missing answers	−.872**	−.086*	

Note. Above diagonal: cohort 1933/1934; below diagonal: cohort 2006.
* p = .01.
** p < .001.

Table 4
Odds ratios (OR) of wrong, right and missing answers (odds 2006/odds 1934).

		Right answers	Wrong answers	Missing answers
Arithmetical Reasoning (A1)	Median	1.26	1.3	0.59
	Mean	1.56	1.39	0.69
	SD	0.85	0.79	0.31
Sentence Completion (A2)	Median	2.64	1.2	0.2
	Mean	3.03	2.33	0.48
	SD	2.32	2.82	0.81
Same–Different (A4)	Median	4.07	0.89	0.05
	Mean	5.5	1.51	0.06
	SD	4.51	1.69	0.08
Computation (B1)	Median	0.94	0.93	2.03
	Mean	1.48	1.18	2.57
	SD	1.73	1	2.27
Information (B2)	Median	1.45	1.96	0.32
	Mean	3.02	3.49	0.83
	SD	5.33	4.83	1.62
Vocabulary (B3)	Median	1.93	1.21	0.22
	Mean	2.72	2.62	0.6
	SD	2.92	4.64	1
Analogies (B4)	Median	3.55	1.02	0.14
	Mean	4.57	1.62	0.26
	SD	4.67	1.89	0.46
Comparisons (B5)	Median	5.92	1.08	0.1
	Mean	8.68	3.18	0.2
	SD	12.25	5.59	0.22

$p = .14$ to $r = .731$, $p < .01$; the mean correlation $r = .426$). More clearly the same relationship is described by the correlations of OR of wrong answers to items with their presentation order ($r = .218$, $p < .36$ to $r = .824$, $p < .01$; the mean correlation $r = .524$). The ORs of right and missing answers to items and their presentation order do not have such a clear and univocal pattern. The odds of giving a wrong answer at the end of a subtest were significantly higher in 2006 than they were in 1933/36.

4. Discussion

We compared the IQ test scores of Estonian student samples from 1933/36 to 2006. Our analysis supported the general finding that IQ scores have risen between comparable age-cohorts in the last decades (Flynn, 2012). The rise in scores is not equal in different subtests. Earlier studies (Must et al., 2003, 2009; Wicherts et al., 2004) showed that the rise in the scores of NIT subtests over time is not correlated with their *g*-loadings, and the comparison of two Estonian student cohorts' (1933/36 and 2006) IQ scores is problematic due to the lack of invariance of measurements. The current analysis allows us to add some new ideas in this series of studies. Estonian FE comparisons are made using data based on the Estonian version of NIT (Tork, 1940). The uniqueness of this test is that the testing procedure and the details of scoring may have an impact on the evaluation of the FE.

In the current analysis we use data at the item level. This means that in addition to summated test scores the answers on the item level are used. The right, wrong and missing answers are separated and analyzed as relatively independent empirical indicators of FE. The NIT testing procedure has several details that may have a significant impact on the results. All the subtests are speeded: the testing time limitations are explicitly stressed by the 20 timing sections. The missing answers seem to be due to these limitations. The

Table 5
The correlation of item presentation order in subtests with the odds and odd ratios (ORs) of different types of answers to the item (Spearman's rank correlation).

	Odds of wrong answers 1934	Odds of wrong answers 2006	OR of wrong answers	OR of right answers	OR of missing answers
Arithmetical Reasoning (A1)	.209	.606*	.824 **	.097	−.068
Sentence Completion (A2)	.393	.580*	.218	.522*	−.532*
Same–Different (A4)	−.096	.402*	.611**	.444**	.444**
Computation (B1)	−.114	.190*	.470*	−.266	−.042
Information (B2)	.335*	.354*	.331*	−.051	.070
Vocabulary (B3)	.705**	.731**	.418**	.220	−.280
Analogies (B4)	−.438*	.267	.747**	.245	.008
Comparisons (B5)	−.304*	.276*	.572**	.580**	−.186
Mean	.086	.426	.524	.224	−.073

* $p < .05$.
** $p < .01$.

time limits per item in different subtests varied more than 12 times (from 1.5 s per item up to 18.8 s per item). The actual time limit per item does not directly reflect the difficulty of the item for students in the sense of how much time the items demand from testees, but rather how much time they are allowed by the test manual.

4.1. The decline in missing answers

Our most important finding is that there have been several trends in test-taking patterns over time. One is the diminishment of the number of missing answers during the period 1933/36–2006 (the mean effect per subtest is about minus 1 *d*). The number of missing answers rose in only one subtest (B1) — the one which required pen and paper calculations (effect = .36 *d*). If we take into account that the main content of the items in this subtest was related to 4 basic arithmetical operations and the 2006 cohort was educated in school for 2 years more than the 1933/36 cohort, then it can be logically assumed that the content of these items was not too difficult for the test-takers. There could be a more complex reason for the missing answers. The subtest had a noticeable time restraint: 10.9 s per item (Table 1). This means that originally the subtest was assumed to require a greater effort than other subtests. The format of this subtest assumed constructed responses (writing results of calculations). There is some empirical evidence that several technical characteristics of the items, including their length and type of response format (multiple choice vs. constructed response), may have had a negative impact on the motivation of those taking the test, especially so in low-stakes testing situations (DeMars, 2000; DeMars & Wise, 2010; Wise et al., 2009). Our interpretation is the following: the exceptional rise in missing answers is due to not the difficulty of the subtest, but rather the avoidance of items that require a great deal of work. The avoidance is understandable as the NIT was administered as a test without consequences for the test-takers.

4.2. The rise of right and wrong answers

The diminishing number of missing answers does not necessarily correspond to a direct rise in right answers. The mean effect of the rise in right answers is .86 *d* per subtest (Table 2). In two subtests (Computation and Information) there was not a rise but rather a fall in number correct. The highest rises of correct answers occurred in the Symbol–Digit (A5) and Comparisons (B5) subtests (1.65 and 1.71 *d* respectively). According to the original test administration manual (Haggerty et al., 1920; Tork, 1940) those subtests allowed a minimal time for completion — respectively 1.5 and 2.4 s per item (Table 1). These small time allowances show that those subtests were not designed to require complicated operations, but mainly perceptual speed. The Symbol–Digit and Comparisons subtests are the analogs to the old Army Beta tests which were designed for the evaluation of the mental capacities of illiterate persons (Yoakum & Yerkes, 1920). Must et al. (2009) showed that those two subtests have the smallest *g*-loadings in the NIT battery (loadings on the first principal component (PC1) in 1933/36 as in 2006, are in the range of .48 to .57). The high rise of subtest scores and their low *g*-loadings contradicts the notion that *g* can account for the gain in these subtests.

Simultaneously with the diminishing number of missing answers and the rise of right answers, there is a concomitant rise in the number of wrong answers (mean effect per subtest .30 *d*). The fact that together with the rise of correct solutions, the number of wrong answers may rise as well, has not been mentioned before in FE literature. From the 9 subtests only one, the Symbol–Digit (A5), showed a minimal negative effect (−.07 *d*), and in all other subtests the number of wrong answers rose. Subtests of Information (B2; 1.10 *d*) and Vocabulary (B3; .67 *d*) showed the greatest increase. These subtests had relatively high *g*-loadings in 1933/36 as well as in 2006 (loadings on the PC1 are in the range from .62 to .84).

4.3. Adjusted scores

The speed of the test-taking was accompanied by a rise in the number of mistakes in 2006. It may indicate that the test-taking pattern of the period 1933/36 to 2006 has changed. Probably the role of guessing in the completion of tests has risen over time (Brand, 1996; Wicherts et al., 2004). If we are to take the position that guessing is involved in test-taking behavior, it is logical to assume that some right answers are also random by chance. A simple and transparent, although very conservative, principle for the adjustment for answers given by chance is to subtract from the number of right answers the number of wrong answers (Fig. 1) as was the intended method for the scoring of the original historical army test (Yoakum & Yerkes, 1920). The effect of this adjustment varies over subtests. The most prominent was the adjustment of the subtest of Arithmetical Reasoning (A1): the test score rise effect dropped from .54 to .15 *d* after the adjustment. There is no reason to believe that the level of students' arithmetical skills has declined and in 2006 they were less adept at finding the right solutions than they were in 1933/36. The widespread use of calculators may have influenced the speed and quality of pen and paper calculations and therefore the rise in wrong answers can partly be attributed to technological changes. In this context it is important to take into account that this subtest has the highest time allowance — 18.8 s per item and its score was weighted by 2 for a full score calculation (Table 2). This means that these items were perceived as time and effort consuming and that the students' results in this subtest are important for the calculation of the complete IQ score. Those arguments support the interpretation that the numerous wrong answers in this subtest are a sign of hurrying, guessing or superficial answering.

Similarly it is possible to interpret the effect of the adjustment on the results of the Sentence Completion (A2) subtest. The subtest gave a relatively generous time limit (12 s per item) and also the weight 2 for a full score was used. The effect of the adjustment is substantial: the recalculated actual right answers diminished the rising effect from 1.24 to .76 *d* after the adjustment. The effect was diminished in some other subtests too after adjustments, and with roughly the same proportions, for example, in the Vocabulary subtest (B3; from .74 to .30 *d*) and the Analogies (B4; from 1.09 to .63 *d*). In the last two subtests the time per item was not considerable (accordingly 6 and 4.5 s per item). In both subtests the original scoring system assigned penalties for wrong answers. This means that the test construction took into account the possibility of guessing. The adjustment for guessing did not have significant consequences on the results of the two subtests which were intended to measure response speed: Symbol–Digit (A5; effects accordingly 1.65 and 1.62 *d*) and Comparisons (B5; effects accordingly 1.71 and 1.61 *d*).

4.4. Odds ratios

The same result — the simultaneous rise of right and wrong answers in test performance was also evident at the item level. It is possible to describe individually all three response types for every item by recourse to odds and FE as odds ratios. This separation allows us to independently investigate the changes in the odds of every item response type across different subtests. The odds ratio is an effect-size statistic for binary data, and is especially appropriate in our case. The change of the item response patterns over time is revealed in the changes of odds. We found that the average OR was rather different from the median ORs. The reason for this discrepancy is the small number of items in the subtests and some extraordinarily high OR values. Thus we used the median as a more appropriate statistic for the description of a central tendency in the changes of OR over time. This analysis supported the findings with subtest averages: the OR of missing answers is smaller than 1 as a rule. The extreme case here was the subtest of Same–Different (A4) in which the missing answers' OR value is only .05. This low OR value means that there are considerably fewer missing answers in this subtest.

The medians of ORs of wrong and right answers, as a rule, are higher than 1. A different analysis – the comparisons of subtest means, as well as the comparisons of medians of ORs – rendered the same conclusion; that as a general trend the rise of right responses to the NIT test items parallels the rise of wrong responses. The response to the item is related to its order in the subtest and this relationship has changed over time. The 2006 cohort tried to solve more items, but in a more haphazard way than did the cohort of 1933/36. The main difference between cohorts is the test-taking speed. But speed has its price — the more items that students tried to solve, the higher the probability of answering incorrectly as well (Table 5). The rank correlation between item order on a subtest and the OR of wrong answers to the items varies from .218 to .814. In contrast, the item order in a subtest does not have any correlation with the OR of missing answers (the mean correlation = −. 073). There is a trend over time to give answers more quickly in spite of the possibility of making errors. Errors are encouraged by the amount of work the items require and by the items' sequence in the subtest. The items in NIT subtests are organized according to an increasing scale of difficulty (Whipple, 1921). Therefore the attempt to answer more items means a higher probability of wrong answers.

4.5. The relationships of answers

On a more general level we found that the NIT is very sensitive to test-taking speed. The main task in completing the NIT items is not to avoid errors but to fill in as much as possible. In most NIT subtests the opposite of a right answer is a missing answer. The NIT is not difficult nor does it facilitate errors, the main question is how many items the student can fill out. In 1933/36 the correlation between the number of correct and missing answers was remarkably high: r = −.959, p < .001 (Table 3). The number of correct answers was not correlated with the number of wrong answers (r = .005). In 2006 the relationship became more complicated: the opposite of a right answer is a missing answer as in 1933/36, but the relationship is weaker (r = −.872, p < .001). A new aspect in 2006 emerges, the correlation between the number of right and wrong answers. The correlation is moderately negative (r = −.367, p < .001). The numerous errors made by respondents in 2006 can have multiple interpretations, such as being indicators of lesser abilities or low motivation. According to classical test theory, every empirical test score consists of two components: the true score and random error. Therefore a lack of correlation between the number of right and wrong answers is not unexpected. The correlation shows that there are other systematical influences (in addition to true score and random errors) that are involved in the measurement.

4.6. An explanation offered

Our findings are best explained by the framework of response time effort and mental taxation categories of the items (Wise & Kong, 2005; Wise et al., 2009; Wolf et al., 1995). In low-stakes testing the test-takers' performance effort will be crucial. Low-stakes conditions facilitate rapid-guessing. The results of rapid guessing are errors, especially so if the items required attention and thought or the test-takers are hurrying towards the end of test. Our analysis indicated that the change in response patterns is parallel to a rise in test scores. Although in 1933/36 as in 2006 the test had no personal consequences for the students, the influence of low-stakes conditions varied.

Test-taking strategies may introduce a construct-irrelevant variance into the test score distribution (Haladyna & Downing, 2004). This means that the IQ test may measure more than latent cognitive ability alone, as has been assumed, because test-taking strategies are also a factor. The same concept can be found in works regarding differential item functioning. Various rates of rapid guessing between groups can manifest detectable levels of item functioning in situations where the item parameters are the same for both groups. Differences between the groups emerge due to the differences in the ways the test is taken. This has an additional impact on the item content in the testing results (DeMars & Wise, 2010). This construct-irrelevant variance and differential item functioning, which is due to the differences in test-taking behavior, may be one reason why measurement invariance in FE studies is problematic. One possible source for the additional attitudinal systematical IQ test score variation is suggested by Brigham (1923, p. 123–126). He explains that the differences in test results are due to a person's adjustment to American society. He found that the IQ score of foreign-born Caucasians who lived in America less than 5 years was approximately 1 SD less than the immigrants who lived in the United States over 20 years, which were at par with Caucasians born in America. Brigham used the terms “hurry-up attitude” and “typically American” for describing the test-taking system for the American Army Mental Tests (alpha and beta) which stresses speed in test-taking. Brigham agreed that adjusting to test conditions is part of an intelligence test, and that this adjustment may be independent of intelligence. Our findings align perfectly with Brigham's ideas — the development of test-taking with a "hurry-up attitude" may partly explain FE.

Our analysis gives empirical support to the previous hypothesis of the relationship of FE to test-taking behavior (Brand, 1996; Wicherts et al., 2004). The change in test-taking speed may be valid. Seventy years ago in 1933/36 there were no real decisions made (for example, admission to schools or employment) on the basis of standardized tests or on the basis of the average school grades in Estonia. Currently test-taking is neither the exception, nor a surprise as there are many tests and examinations which students need to pass. In a modern society grade point averages (GPAs) or the results of tests are important in various real life (selection) situations. It is perhaps an implicit rule now to be fast, and to do as much as possible. At the same time, over the decades there has been less emphasis on being correct and trying not to make mistakes. Frequently, wrong answers have no serious consequences for test takers. On the other hand, it is important to take into account that the philosophy of the NIT is strongly influenced by the application of mental testing in a military context. The NIT assumes order and full compliance with the testing requirements. Over time the schoolchildren's obedience to the testing rules in a group-testing situation has evidently relaxed. Tests that do not have important consequences for the test-taker are not taken seriously. From the viewpoint of psychometrics it is clear that in the 1920s guessing in test-taking was possible, and some right answers may be the result of guessing (Yoakum & Yerkes, 1920, p. 6–7). The same is true now. The changes in test-taking guessing over time have direct consequences on the estimation of the FE.

4.7. A metacognitive perspective

Research of learning efficiency has shown that metacognition is an important predictor of the learning performance of students (e.g. Wang, Haertel, & Walberg, 1990; van der Stel & Veenman, 2010). Metacognitive skills concern the knowledge that is required for the control one's learning activities — metacognition refers to a learner's awareness about learning. There is research evidence that intellectual and metacognitive abilities are positively correlated, but still with different skills (van der Stel & Veenman, 2010). Veenman and Spaans (2005) argue that metacognitive skillfulness outweighs intelligence as a predictor of learning performance and that metacognitive skills appear to be general for third-year students, but rather domain-specific for first-year students. It is possible to incorporate this framework into the test-taking process also. Students in modern societies have broad experience with different testing and examination procedures and their consequences. It is logical to assume that they are able to manage their test-taking resources just as they are able to manage different learning processes. Undemanding test-taking may be the right strategy for a low-stakes examination situation and the FE may partly reflect the development of this skill over time. "When our scales measure the non-intellective as well intellectual factors in intelligence, they measure what in actual life corresponds to intellectual behavior" (Wechsler, 1943, p.103).

5. Conclusion

FE as a research paradigm was inaugurated decades ago in the 1980s (Lynn, 1982; Flynn, 1984, 1987). At the time it was not considered that other factors such as changes in test taking patterns might be behind the IQ test score rise over time. The simultaneous influence of power and speed on the test results is explained by our finding that the relationship between right, wrong and missing answers varies across subtests. There is no single best option to eliminate the influence of negative aspects of test-taking (e.g. guessing) on test results due to there being several variables involved: different subtests can measure different latent cognitive abilities; there may be a different number of items and answer options, each test may have different time limits, or the item characteristics (more or less difficult, more or less discriminating) may vary. The next step would be to model this interplay more exactly and in more detail at the level of latent variables.

Juhan Tork (1889, Tartu, Estonia — 1980, Toronto, Canada). An Estonian educator and psychologist. In the early 1930s Tork made an adaptation of the National Intelligence Tests for Estonia. He tested about 6000 schoolchildren and calculated the Estonian national IQ norms. His doctoral dissertation was titled "The Intelligence of Estonian Children" (Tork, 1940).

Acknowledgments

A Primus grant (#3-8.2/60) from the European Social Fund to Anu Realo supported the preparation of this article. We thank Jelte Wicherts for helpful comments on an earlier version of this article.

References

Atkinson, J. (1957). Motivational determinants of risk-taking behavior. *Psychological Review*, *64*, 359–372.

Barry, C. L. (2010). Do examinees have similar test-taking effort? A high-stakes question for low-stakes testing. *International Journal of Testing*, *10*, 342–363.

Bates, T., & Eysenck, H. (1993). Intelligence, inspection time, and decision time. *Intelligence*, *17*, 523–531.

Baumert, J., & Demmrich, A. (2001). Test motivation in the assessment of student skills: The effects of incentives on motivation and performance. *European Journal of Psychology of Education*, *16*, 441–462.

Brand, C. (1996). *The g Factor — General Intelligence and Its Implications.* John Wiley & Sons Inc.

Brigham, C. (1923). *A Study of American Intelligence.* Princeton: Princeton University Press.

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

Deary, I., & Stough, C. (1996). Intelligence and inspection time. *American Psychologist*, *51*, 599–608.

DeMars, C. (2000). Test stakes and item format interactions. *Applied Measurement in Education*, *13*, 55–77.

DeMars, C., & Wise, S. (2010). Can differential rapid-guessing behavior lead to the differential item functioning? *International Journal of Testing*, *10*, 207–229.

Duckworth, A. L., Quinn, P. D., Lynam, D. R., Loeber, R., & Stouthamer-Loeber, M. (2011). Role of test motivation in intelligence testing. *Proceedings of the National Academy of Sciences of the United States of America*, *108*, 7716–7720.

Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology*, *53*, 109–132.

Eklöf, H. (2010). Skill and will: Test-taking motivation and assessment quality. *Assessment in Education: Principles, Policy & Practice*, *17*, 345–356.

Flynn, J. (2012). *Are We Getting Smarter? Rising IQ in the Twenty-first Century.* Cambridge: Cambridge University Press.

Flynn, J. (1984). The mean IQ of Americans: Massive gains 1932 to 1978. *Psychological Bulletin*, *95*, 29–51.

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

Furneaux, W. D. (1960). Intellectual abilities and problem solving behaviour. In H. J. Eysenck (Ed.), *Handbook of Abnormal Psychology*. (pp. 167–192) London: Pergamon Press.

Haggerty, M., Terman, L., Thorndike, E., Whipple, G., & Yerkes, R. (1920). *National Intelligence Tests. Manual of Directions. For Use with Scale A, Form 1 and Scale B, Form 1.* New York: World Book Company.

Haggerty, M., Terman, L., Thorndike, E., Whipple, G., & Yerkes, R. (n.d.-a). *National Intelligence Tests. Scale A - Form2*. London: George G. Harper & CO.

Haggerty, M., Terman, L., Thorndike, E., Whipple, G., & Yerkes, R. (n.d.-b). *National Intelligence Tests. Scale B - Form2*. London: George G. Harrap & CO.

Haladyna, T., & Downing, S. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice*, *23*, 17–27.

Jensen, A. (1980). *Bias in Mental Testing.* New York: Free Press.

Jensen, A. (1998). *The g Factor. The Science of Mental Ability.* Westport: Praeger.

Klein Entink, R., Kuhn, J., Hornke, L., & Fox, J. (2009). Evaluating cognitive theory: A joint modeling approach using responses and response times. *Psychological Methods*, *14*, 54–75.

Lee, Y., & Chen, H. (2011). A review of recent response-time analyses in educational testing. *Psychological Test and Assessment Modeling*, *53*, 359–379.

Lynn, R. (1982). IQ in Japan and the United States shows a growing disparity. *Nature*, *297*, 222–223.

Lynn, R., & Harvey, J. (2008). The decline of the world's IQ. *Intelligence*, *36*, 112–120.

Must, O., te Nijenhuis, J., Must, A., & van Vianenen, A. (2009). Comparability of IQ scores over time. *Intelligence*, *37*, 25–33.

Must, O., Must, A., & Raudik, V. (2003). The secular rise in IQs: In Estonia, the Flynn effect is not a Jensen effect. *Intelligence*, *31*(5), 461–471.

Partchev, I., De Boeck, P., & Steyer, R. (2011). How much power and speed is measured in this test? *Assessment*, *20*, 242–252.

Pintrich, P., Cross, D., Kozma, R., & McKeachie, R. (1986). Instructional psychology. *Annual Review of Psychology*, *37*, 611–615.

Rindler, S. (1979). Pitfalls in assessing test speededness. *Journal of Educational Measurement*, *16*, 261–270.

Rushton, J. (1999). Secular gains in IQ not related to the g factor and inbreeding depression—Unlike Black–White differences: A reply to Flynn. *Personality and Individual Differences*, *26*, 381–389.

Sundre, D. L., & Kitsantas, A. (2004). An exploration of the psychology of the examinee: Can examinee self-regulation and test-taking motivation predict consequential and non-consequential test performance? *Contemporary Educational Psychology*, *29*, 6–26.

te Nijenhuis, J., & van der Flier, H. (2013). Is the Flynn effect on g?: A meta-analysis. *Intelligence*, *41*, 802–807.

Terman, L. (1921). *The Intelligence of School Children.* London: George G. Harrap.

Tork, J. (1940). *Eesti laste intelligents.* Tartu: Koolivara.

van der Linden, W. (2011). Test design and speededness. *Journal of Educational Measurement*, *48*, 44–60.

van der Maas, H., Molneaar, D., Maris, G., Kievit, R., & Borsboom, D. (2011). Cognitive psychology meets psychometric theory: On the relation between process models for decision making and latent variable models for individual differences. *Psychological Review*, *118*, 339–356.

van der Stel, M., & Veenman, M. (2010). Development of metacognitive skillfulness: A longitudinal study. *Learning and Individual Differences*, *20*, 220–224.

Veenman, M. V. J., & Spaans, M. A. (2005). Relation between intellectual and metacognitive skills: Age and task differences. *Learning and Individual Differences*, *15*(2), 159–176.

Wang, M., Haertel, G., & Walberg, H. (1990). What influences learning? A content analysis of review literature. *The Journal of Educational Research*, *84*, 30–43.

Wechsler, D. (1943). Non-intellectual factors in general intelligence. *Journal of Abnormal and Social Psychology*, *38*, 101–103.

Whipple, G. (1921). The National Intelligence Tests. *The Journal of Educational Research*, *4*, 16–31.

Wicherts, J. M., & Dolan, C. V. (2010). Measurement invariance in confirmatory factor analysis: An illustration using IQ test performance of minorities. *Educational Measurement: Issues & Practice*, *29*(3), 39–47.

Wicherts, J. M., Dolan, C. V., & Hessen, D. J. (2005). Stereotype threat and group differences in test performance: A question of measurement invariance. *Journal of Personality and Social Psychology*, *89*, 696–716.

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., & Zand Scholten, A. (2010). Test anxiety and the validity of cognitive tests: A confirmatory factor analysis perspective and some empirical findings. *Intelligence*, *38*, 169–178.

Wise, S., & DeMars, C. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, *10*, 1–17.

Wise, S., & DeMars, C. (2010). Examinee noneffort and the validity of Program Assessment Results. *Educational Assessment*, *15*, 27–41.

Wise, S., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*, 163–183.

Wise, S., Pastor, D., & Kong, X. (2009). Correlates of rapid-guessing behavior in low-stakes testing: Implications for test development and measurement practice. *Applied Measurement in Education*, *22*, 185–205.

Wolf, L., & Smith, J. (1995). The consequence of consequence: Motivation, anxiety, and test performance. *Applied Measurement in Education*, *8*(3), 227–242.

Wolf, L., Smith, J., & Birnbaum, M. (1995). Consequence of Performance, Test Motivation, and Mentally Taxing Items. *Applied Measurement in Education*, *8*, 341–351.

Yoakum, C., & Yerkes, R. (1920). *Army Mental Tests.* New York: Henry Holt and Company.