Journal of open psychology data

Must, O and Must, A 2014 Data from "Changes in test-taking patterns over time" concerning the Flynn Effect in Estonia. *Journal of Open Psychology Data,* 2(1): e2, DOI: http://dx.doi.org/10.5334/jopd.ab

DATA PAPER

Data from "Changes in test-taking patterns over time" concerning the Flynn Effect in Estonia

Olev Must,[1] Aasa Must[2]

[1] Department of Psychology, University of Tartu, Tartu, Estonia

[2] Estonian National Defence College, Tartu, Estonia

The dataset from our previous *Intelligence* paper [1] consists of data collected from the National Intelligence Tests (NIT, Estonian adaptation) at two historical time points: in 1933/36 (N=890) and in 2006, (N=913). The average age of the students was 13 years. The data-file consists of information about the cohort, age, and gender and test results at the item level for nine of the ten NIT subtests and subtest scores for the 10th subtest. The three answer types are separated into three categories: right answers, wrong answers and missing answers. Data can be used for the psychometric research of cohort and sex differences at the scale and item level.

Keywords: Flynn Effect; National Intelligence Tests; Tork; test taking pattern; Estonia

Funding Statement: Estonian Scientific Foundation: grant no 2387 and 5856. European Social Fund: a Primus grant (#3-8.2/60) to Anu Realo. Baylor University financial support for data quality control in the archive.

(1) Overview

Context

The study concerns the Flynn Effect (FE), or the secular increase in intelligence (IQ) test scores over time. Two cohorts of the same age are compared based on the results of the same 10 subtests of mental abilities. Must and Must [1] provide a detailed account of the study and the data collection from the 1930s and from 2006.

Two unique historical circumstances made this study possible. In the early 1930s Estonian educators sought instruments which would allow them to measure a student's intelligence to be able to influence their educational careers for the better. It was decided to adapt the National Intelligence Tests into Estonian – tests that were originally developed to measure students' IQ in United States [2, 3, 4, 5]. This work was headed by Juhan Tork. He took this duty quite seriously as the analysis of the test adaptation process and the elaboration of IQ norms of Estonian students were the main focus of his PhD dissertation: "The Intelligence of Estonian Children" in 1939 [6].

Although The Second World War and Soviet occupation set back Estonian society and intellectual life for half of a century, most of the completed test materials were kept in the repository of the Estonian National Historical Archive. Although most of the data was collected from southern Estonia, Tork's interpretation was that the results of his work were representative for entire Estonia. The availability of real historical test- data made the Flynn Effect comparisons possible decades later. The second measurement, undertaken with the aim to estimate the FE in Estonia was done in 2006.

Collection date(s)

Estimation of the FE means comparisons of test results of comparable cohorts from different time periods. The first testing was done in 1933 - 1936 and the second in 2006.

Background

The scientific roots of this project are related to the history of IQ measurement. The development and application of IQ tests in the army recruiting process during The First World War was successful and led up to a major expansion of the use of IQ tests beyond the army. The American army tests (Army Alfa and Army Beta) served as the basis for the development of the next generations of testing instruments and approaches. The test, titled *National Intelligence Tests* (NIT) sought to measure the intelligence of schoolchildren (grades 3 to 8) and was developed 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 [7]. The NIT was adapted into Estonian at the beginning of the 1930s. The test name in Estonian is "Intelligentsustestid" (Intelligence tests).

(2) Methods

The NIT is a timed test administered in paper-and-pencil format. It consists of two complementary scales (A and B) in two parallel forms (I and II). Both scales – A and B – consist of 5 subtests.

- Arithmetical Reasoning (A1). The subtest consists of 16 items that require the test taker to find a solution

to an unknown quantity. For example: "How many seats are there in 7 rooms, if each room has 30 seats?"
- Sentence Completion (A2). The subtest consists of 20 items that require the test taker to fill in a missing word to make sentence understandable and correct. An example: "The letter …… came …… good news".
- Concepts (A3). The subtest consists of 24 items requiring the selection of two characteristic features from among those given. For example, "apple: basket/redness/seeds/skin/sweetness".
- Same–Different (A4). The subtest consists of 40 items requiring the evaluation of whether the meaning of the words presented is the same or different. For example, "brief….short; elevate…raise".
- Symbol–Digit (A5). The subtest consists of 120 items requiring a decision as to which digit should be assigned a symbol based on a key; 9 different symbols were presented. An abridged example of the key: ∞ □ Δ 1 2 3
- Computation (B1). The subtest consists of 22 items requiring addition, subtraction, multiplication, and division of both integers and fractions. For example, subtract: 38260 – 17700.
- Information (B2). The subtest consists of 40 items of everyday knowledge. For example, "The day before Thursday is: Wednesday/Tuesday/Friday/Monday".
- Vocabulary (B3). The test consists of 40 items requiring knowledge of the qualities of different objects (*yes* and *no* answers), for example, "Does a dog have five legs?"
- Analogies (B4). The test consists of 32 items requiring the transference of the relation of two given words to other presented words. For example, baby – cries;
- cat - : mews/hole/little/dog.
- Comparisons (B5). The test consists of 50 items requiring judgment of the sameness of sets of numbers, family names, and graphic symbols presented in two columns. For example, Lindpere A. J. … Lindpere J. A.

Both the original NIT-tests in English and the Estonian adaptation are subject to copyright and are available from the authors upon request.

Samples

The data for the sample from 1933/36 was taken from the Estonian National Historical Archive (foundation EAA.2101). There is data from approximately 5000 persons in the archive depository. However various persons have completed different combinations of test forms and scales (for example, scale A form I and scale B form II etc.).

The main inclusion criterion in the sample formation for the current study was that all students in this sample have filled out both forms of the second version of the test. Our preliminary pilot study revealed that the second version of the scales (A II and B II) is more appropriate for re-use decades later (due to the content of vocabulary and information subtest). Hence we sought only those specific cases from the archive that met this inclusion criterion.

The second inclusion criterion was the school grade. In the 1930s, the compulsory basic school had 6 grades. The intellectual level of graduates of this educational level was of great interest to J.Tork. At the same time he tested students from other grades also, mainly those preceding the 6th grade. The largest portion of historical data derives from students of the 4th to 6th grades. The typical age of this group is 12 to 14 years. Those parameters guided the selection of a younger cohort in our study. In 2006 this group corresponded mostly with students from the 6th to 8th grades.

The formation of the sample of the younger cohort was guided by the aim of being as comparable as possible to the older one. This means that we were seeking schools and students from the same region (mainly from southern Estonia), and from the grades 6 to 8, where the typical student is 12 to 14 years old.

The older cohort (1933/36; N = 890) consisted of students from grades 4 to 6, with a mean age of 13.3 (SD = 1.24) years whereas the younger cohort (2006, N = 913) were from the grades 6 to 8 with a mean age of 13.5 (SD = .93) years.

The samples are different at least in two important points:

- Educational differences. The younger cohort has had two more years of education than the older one. This difference is related to the lowering of the age of obligatory school attendance in Estonia. The educational difference may have a significant impact on the test results. For Tork himself schooling differences were also important. He found that the students' age and grade were intertwined with the test results. Tork [6, p 192- 205] devoted a whole chapter to this co-variation and calculated age-grade IQ norms. The same-aged students in the higher grades averaged higher test scores.
- Urbanization differences. Tork [6, p 212-218] noticed that the rural-urban dimension creates significant differences in test results. He found that this difference is equal to one year of schooling. He believed that there should be different IQ norms for urban and rural students. Changes over the decades in the urbanization rate of the population may have influenced the estimation of the FE in Estonia.

In both cohorts the tests were administered in groups, based on school-grade, in a regular classroom, during a regular school day. Students who were absent from school on the testing day were not tested. There were only a few absentees in each group. We interpreted this as a random event. The participation rate was not fixed.

Procedures

The testing procedures of the NIT resembled testing in a military context. From test-takers, full obedience to the procedures was required. "Directions, and especially commands, should be spoken authoritatively, and instant obedience should be expected and required. Every child

	Pre-test		Test			
	No of items in pre-test	Minutes per pre-test	No of items in subtest	Minutes per subtest	Sec per item	Scoring algorithm
Arithmetical Reasoning (A1)	6	1	16	5	18.8	Number of right answers × 2
Sentence completion (A2)	10	0.5	20	4	12	Number of right of answers × 2
Concepts (A3)	8	0.5	24	3	7.5	Partial credit system (1 and 2 points for right answers); sum of credit points
Same – Different (A4)	20	0.5	40	2	3	Number of right answers – number of wrong answers
Symbol – Digit (A5)	20	0.5	120	3	1.5	Number of right answers × 3/10
Computation (B1)	10	1	22	4	10.9	Number of right of answers × 2
Information (B2)	16	0.5	40	4	6	Number of right of answers
Vocabulary (B3)	15	0.5	40	3	4.5	Number of right answers – number of wrong answers
Analogies (B4)	12	0.5	32	3	5.6	Number of right of answers
Comparisons (B5)	20	0.5	50	2	2.4	Number of right answers – number of wrong answers

Table 1: Time limits and the NIT scoring system

should obey promptly and without question." [3, p.6; 6, p.69]. Exercise examples of items preceded each subtest. The procedure established exact timing requirements which were set for both pre-subtests and real tests (**Table 1**). The NIT procedure included 20 different timing episodes. Commands such as "Ready — Go!", "Stop!", "Pencil up!" are typical commands in the NIT testing manual. [6, p 69 – 75]).

A and B scales were printed and administered as independent tests. Both booklets included the title page for a person's background information. Information on the title page was used to join data from scales A and B into one record. As a rule, the A scale was used first and the B after a little break. In several cases the testing took place over several days.

Quality control

The scoring and data computerization were checked twice. We also used logic analysis (possible values of computerized data) and checked for extreme values (age). It was possible to check the declared age (years and months) by comparing the birth and testing date.

In both cohorts there were some test booklets that were only superficially completed (sporadic answers in one or several subtests); these results were not included in the data file. The number of such test booklets was very small (<10). We interpreted this as an indicator of a lack of any motivation to do well on the test.

Ethical issues

The testing of both cohorts was done in accordance with the local guidelines of student examination and testing. In 1933/36 the testing was supported by the Estonian Ministry of Education. There is no evidence that some schools did not allow students to take the test. Evidently one of the main criteria of school selection was the location of the school, this was due to the transport cost and time resources. The majority of schools were located in the same region (southern Estonia) as the researcher's office.

In 2006 the decision to allow testing was made at the school level. Some school authorities refused to take part as the testing disturbs educational process. The testing was done during regular school hours in regular classrooms. Actually, since testing took place during an ordinary school-day all the students were asked to take part. The participation was voluntary for students (both in 1934/36 and in 2006) in the sense that the test-takers were not forced to take the test, and the test had no consequences for the test-takers. In both cohorts there were some students (about 1%) who filled out the tests in a very superficial manner (occasional answers). We interpret this as a formal conformity to take the test without any real interest in doing well on it. Test-booklets with only random answers were not included in the data-file.

The testing was not anonymous – test-takers had to write their names and background data on the first sheet of the test booklet (school, grade, their parents' employment, the number of children in the family and their birth order, the parents' birth place, age, ethnicity). The Estonian adaptation required more personal information from students than the original NIT. Presenting personal information in testing and school examinations was a relatively typical practice in Estonia. In some cases, in the test booklets, the corresponding fields about personal information on the test-booklets were left empty, or filled in very generally.

Estonian regulations do not require additional parental approval for testing students in schools. All data is sufficiently anonymised to prevent identification of individual students (based on the information in the data files).

The completed historical test booklets are in the public archive. Personal information from the 2006 cohort is not publicly available and not used in any way. The test-takers, their parents and teachers were not given individual feedback on the test results. In our research we operated under the ethical testing and examination standards accepted by the Estonian educational system.

(3) Dataset description

Object name

Data: FE_1933_2006_data.csv
Description of variables and coding: FE_1933_2006_variables_and_coding.csv

Data type

The file consists of secondary data. This means that direct answers are coded according to a key at the item level (true, wrong, missing). In one subtest (A5) the right, wrong and missing answers were tallied at the test site by the test administrators, from the test booklet.

Format names and versions

The files format: comma separated variables (.csv).

Data collectors

- Vilve Raudik (preparation of the new test layout, testing, data coding).
- Triin Lett (testing, data coding).
- Ene Karja (archival work).
- Aasa Must (data collection, archival research with historical data, data coding).
- Olev Must (data collection, archival research with historical data, data coding).

Language

The test language was Estonian. Data documentation is in English.

License

CC0

Repository location

http://doi.org/10.7910/DVN/23791

Publication date

01 January 2014

(4) Reuse potential

The data set is quite extensive because it includes item-level data from nine of the 10 NIT subtests of two large cohorts. The data may be reused for different purposes: e.g. (1) re-analysis of the results in the *Intelligence* paper, (2) for teaching purposes, and (3) for further analysis (including collaborative projects with Must and Must). The data is useful for advanced psychometric analyses at both the item level (Item Response Theory models) and/or the scale level (Structural Equation Modelling with latent variables) for the study of the Flynn Effect, or sex differences.

Acknowledgements

Vello Paatsi (Estonian Literary Museum, Estonia)
Jelte Wicherts (Tilburg University, The Netherlands)
A. Alexander Beaujean (Baylor University, United States)

References

1. **Must, O** and **Must, A 2013** Changes in test-taking patterns over time. *Intelligence*, 41: 780–790. DOI: http://dx.doi.org/10.1016/j.intell.2013.04.005
2. **Brigham, C** 1923 *A Study of American Intelligence*. Princeton: Princeton University Press.
3. **Haggerty, M, Terman, L, Thorndike, E, Whipple, G** and **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.
4. **Terman, L** 1921 *The Intelligence of School Children*. London: George G. Harrap.
5. **Whipple, G** 1921 The National Intelligence Tests. *The Journal of Educational Research*, 4: 16–31.
6. **Tork, J** 1940 *Eesti laste intelligents*. Tartu: Koolivara.
7. **Yoakum, C,** and **Yerkes, R** 1920 *Army Mental Tests*. New York: Henry Holt and Company.

Peer review comments: http://openpsychologydata.metajnl.com/downloads/peerreview/jopd-ab.pdf

How to cite this article: Must, O and Must, A 2014 Data from "Changes in test-taking patterns over time" concerning the Flynn Effect in Estonia. *Journal of Open Psychology Data,* 2(1): e2, DOI: http://dx.doi.org/10.5334/jopd.ab

Published: 18 March 2014

]u[The *Journal of Open Psychology Data* is a peer-reviewed open access journal published by Ubiquity Press

OPEN ACCESS