Peer Review Comments

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

Article submitted: 9 December 2013

Article accepted: 14 February 2014

Article published: 18 March 2014

Editor: Jelte M. Wicherts

Responses for Version 1

Reviewer A: Wendy Johnson
Review Completed: 9 December 2013

'Must and Must provides a detailed account of the study and data collection. The study concerns the Flynn effect, or the secular increase in IQ scores. Two same-age cohorts are compared on the results of the same 10 subtests of mental abilities.'
Reviewer comment: Some information needs to be provided here on what the purpose of the study was, and how the sample was recruited and tested, including the extent to which it was representative of the population.

'The first wave data were collected in the period 1934 -1936. The second wave was collected in 2006.'
Reviewer comment: Wave' usually refers to longitudinal samples. But clearly that's not what these were.

'At the beginning of 1920s the National Intelligence Tests (NIT) was developed on the basis of Army Alfa and Army Beta tests for testing mental abilities of school children (3-8 grades) [2-5].'
Reviewer comment: Say whose Army tests these were. And Army tests are not for young children typically. How were the tests adapted to be developmentally appropriate, and how was the validity of that tested?

'The NIT is a paper-and-pencil timed test. It consists of two scales (A and B) in two forms. Scale A consists of 5 subtests: Arithmetical Reasoning (A1), Sentence Completion (A2), Logical Selection (A3), Same–Opposite (A4) and Symbol–Digit (A5). Scale B involves the following 5 subtests: Computation (B1), Information (B2), Vocabulary (B3), Analogies (B4) and Comparisons (B5) (see Must & Must, 2013 [1] for more details).'
Reviewer comment: Are the two forms intended to be parallel and interchangeable, or are they intended to be complementary to each other?

'The older cohort (1933/36; N = 890) consisted of students from grades 4 to 6, mean age 13.3 (SD = 1.24) years, and the recent cohort (2006, N = 913) from grades 6 to 8 with mean age 13.5 (SD = .93) years.'
Reviewer comment: So the students were very much the same age, but it looks like they'd had rather different amounts of education. This needs comment and explanation (dates, ages, how universally implemented) and the implications for any assessment of 'true' Flynn effect need to be addressed. Also, the early cohort would be before the Soviet Union and the more recent cohort after. What differences might this make in terms of educational resources and objectives and techniques and who attended school?

'The scoring and data computerization were checked twice. We also used a logical analysis and checked for extreme values (age).'
Reviewer comment: What is a 'logical analysis'? What ages were considered extreme? What was done with those considered extreme and how many of them were there? Why not just adjust for age?

'The testing in 2006 was approved by the administration of the school in accordance with local guidelines.'
Reviewer comment: Do you mean the schools the students attended? How were these schools selected, and who attended them?

'The original NIT as the Estonian adaptation was not anonymous – the test-takers had to write their names and background data on the first sheet of the test booklet.'
Reviewer comment: What sort of data? Would this have introduced stereotype threat? What difference might this make?

'There were some tests fulfilled very formally (occasional answers) in both cohorts; these results are not included in the datafile.'
Reviewer comment: What does this mean? Some test items received only sporadic responses? Why would this have been? Do you mean that these items are not in the data file, or the whole tests in which they are located?

'The testing was administered only in the schools where the administration approved the study.'
Reviewer comment: What proportion of solicited schools was this? How did the cohorts differ on this?

'This means that direct answers are coded according to the key at the item level. Missing item-scores are also given in the data. In one subtest (A5) the right, wrong and missing answers were counted directly by the test administrators from the test booklet.'
Reviewer comment: Not clear what this means.

Reviewer B: Dylan Molenaar
Review Completed: 9 December 2013

I think the data is valuable as:
(1) it contains item level data which is not always the case in intelligence test score data sets (I've seen many data sets on intelligence which only included sum scores or even only IQ scores);
(2) it includes a large number of subjects (890 and 913 = 1803);
(3) it is interesting that two highly separated cohorts are part of these data.

Minor comment: In their rapport, the authors refer to their intelligence paper often, which makes the report not a stand alone document..

At first, the number of missing variables seemed large, but when looking in more depth, I noticed that it are mostly the later items within a subtest that are missing, which is common in intelligence testing as subjects may skip more difficult items. The file including the variable names was ok, in the sense that I could work with it.

It required some effort however as variable labels and answer category labels are intermixed. But I assume that this is something from SPSS (I studied the data using R for which such a file is suboptimal). But my general opinion is that the data is certainly useful.

Responses for Version 2

Reviewer A: Wendy Johnson
Review Completed: 1 January 2014

The paper is much improved. The main thing needed is information on how many/what proportion of students didn't take the test in the younger cohort and how many were thrown away because they gave only sporadic answers. Any such information from the older cohort would also be helpful.

The English needs a lot of polishing.