Intelligence 57 (2016) 7–14

Contents lists available at ScienceDirect

Intelligence

Predicting the Flynn Effect through word abstractness : Results from the National Intelligence Tests support Flynn's explanation

Olev Must *, Aasa Must, Jaan Mikk

Institute of Education, University of Tartu, Salme 1a, 50103 Tartu, Estonia

ARTICLE INFO

Article history:
Received 20 January 2016
Received in revised form 16 March 2016
Accepted 16 March 2016
Available online 2 April 2016

Keywords:
Flynn Effect
National Intelligence Tests
Abstract thinking
Guessing
Tork
Estonia

ABSTRACT

The current study investigates the Flynn Effect (FE) and its relation to abstract thinking ability. We compare two cohorts of Estonian students (1933/36, n = 888; 2006, n = 912) using the Concepts (Logical Selection) subtest of the Estonian adaptation of the National Intelligence Tests (NIT). The item presentation order of the subtest correlates with the abstractness of the words used in the items (r = .609) of the subtest. The different test results (right, wrong and missing answers) were analysed in order to make an estimate of the FE magnitude. The FE for abstract thinking ability of those samples was 1.06 Hedges' g (adjusted for guessing). The magnitude of the FE is dependent upon the degree of difficulty of the items (an item's difficulty is estimated by determining its abstractness and its familiarity to students). The more difficult part of the subtest (the second half) showed a FE = 1.80 whereas the easier part (the first half) of the subtest showed a FE = .72. Word abstractness was a strong predictor of all the testing results in both cohorts (Beta = .700). The familiarity of words used in the test items has no correlation with the test results if word abstractness is controlled in both cohorts. Our findings support Flynn's explanation that the FE is primarily an indicator of the rise in abstract thinking ability.

1. Introduction

The Flynn Effect (FE) is the substantial rise of IQ scores over time (Herrnstein & Murray 1994; Flynn 2007, 2012, 2013). Recently Trahan, Stuebing, Fletcher, & Hiscock (2014) concluded in their meta-analysis of FE research (N = 285 studies since 1951) that the FE is not diminishing in modern societies. They have also concluded that the sample cohorts, and the order in which the test is administered might have a significant impact on future estimations of the FE. The findings demonstrate that FE research will continue to seek new evidence and explanations for the FE phenomenon. Some of these new developments are controversial.

There has been research that presents new data and analyses indicating that decades and centuries ago, human populations were intellectually more productive, and exhibited a faster simple reaction time than modern humans do. These results would then suggest that generations ago humans actually had better information processing abilities, and therefore greater mental abilities (e.g., te Nijenhuis & van der Flier 2013; Woodley, te Nijenhuis, & Murphy 2013; Woodley of Menie et al., 2015). The opposite seems to be true as well. The research of Amstrong et al. (2016) offers new empirical evidence that supports Flynn's (2007, 2012, 2013) position, finding that if confounding variables are controlled (subtest g-loadings and guessing), then the abstractness of the subtests (determined by expert opinion) can positively predict FE magnitudes.

1.1. Different approaches

Historically FE research has been related to g-theory (Jensen 1998; Flynn 2007; Spearman 1927). According to g-theory, people differ in their general ability (g) to solve cognitive problems. A. Jensen (1998) applied the principal component analysis that described a subtests' relatedness to g. The subtests loadings on a common general factor constitute an independent vector, which can then be correlated with other vectors, such as, for example, those of secular IQ gains in test scores. Rushton (1999) was the first to present evidence showing that the secular rise in IQ scores has no correlation with the g-vector. And there are other works as well that confirm Rushton's findings (Must, Must, & Raudik 2003). There are also findings that confirm that the FE is in fact the Jensen Effect (e.g., Colom, Juan-Espinosa, & Garcia, 2001). Technically, for FE research it is actually quite simple to calculate the g-vectors of different samples and compare the loading differences. This approach assumes, that comparable subtests or items have the same meaning (they are invariant).

The widespread conception of g stresses the biological differences between people, but cultural context and its influences are not important factors for describing g. According to Jensen's (1998) g-theory, heritability and environment are independent factors of human development. This framework suggests that brain power is relatively constant and not shaped by environmental influences. Jensen (1998,

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

http://dx.doi.org/10.1016/j.intell.2016.03.003
0160-2896/

p.332) also proposed that the IQ increments should be divided into two parts over time. The functional g-loaded part would indicate the biological improvements that produce the general biological effects. The "hollow" with respect to g part of the IQ increments indicates flexible environmental changes that are not based on true biological factors. Kan, Wicherts, Dolan, & van der Maas (2013) came up with a different approach. They found that the subtests' g-loadings correlate with their relatedness to a culture, or more specifically to its cultural load. The more culturally-loaded a subtest is, the better it is able account for the variance of IQ. This finding indicates that FE research should be more sensitive to cultural and environmental influences, especially in regards to the influence of language and vocabulary.

Instead of simply accepting the independence of genotype and environment, Kan, Wicherts, Dolan, & van der Maas (2013) argued that it is more appropriate to take into account their covariance in the development of IQ. "Because the acquisition of knowledge depends on cognitive processing, individuals who develop relatively high levels of cognitive-processing abilities tend to achieve relatively high levels of knowledge. High achievers are more likely to end up in cognitively demanding environments that encourage and facilitate the further development of a wide range of knowledge and skills." (Kan, Wicherts, Dolan, & van der Maas 2013, p. 2426). Their data shows that certain subtests, such as the Information and Vocabulary (of the WAIS), are highly informative in regard to their high g-load, as well as via their cultural load. This viewpoint stresses the importance of seeking new evidence to show the ways in which culture can contribute to the rise in IQ.

By its nature, an estimation of the FE assumes a comparison of IQ measurements at different timepoints. This comparison assumes, that the measurements have the same meaning (they are invariant). When making comparisons of the FE at different timepoints, however, subtests with a high cultural load tend to show a lack of measurement invariance.

1.2. Measurement invariance

Measurement invariance means that the variables used in the comparisons have the same content and meaning. The measurement is invariant if two people with the same level of a latent trait achieve the same test score (Mellenbergh 1989). Wicherts, Dolan, Hessen, Oosterveld, van Baal, Boomsma, et al. (2004) found that every paper about the FE that they studied violated this principle. This means that IQ test scores had a different meaning for the test-takers at different timepoints. In such a case, the most common solution is to eliminate the variables that are causing the non-invariance, or to free up some of the parameters of the equations. This classical solution for handling the invariance in the language of parameters of structural equations (intercepts, loadings, variances) has been used in several of the FE studies (Must, te Nijenhuis, Must, & van Vianenen 2009; Shiu, Beaujean, Must, te Nienhuis, & Must 2013; Wicherts, Dolan, Hessen, Oosterveld, van Baal, Boomsma, et al. 2004).

2. Abstract thinking

Flynn (2009) himself has found that the FE is more prominent in the results of the Vocabulary and Similarity subtests of the WAIS. He has also posited that the FE is actually an indication of a rise in abstract thinking ability (Flynn 2007, 2012, 2013). These particular subtests are comprised of cultural knowledge, and additionally, the items of these subtests also measure the ability to divide objects into abstract parts, and manipulate them. These are precisely the type of operations that are taught in schools. It was Wicherts, Dolan, Hessen, Oosterveld, van Baal, Boomsma, et al. (2004) who estimated the measurement invariance of 5 of the tests used in various FE research papers and described the subtests that violated the invariance. The tests with high cultural loadings (those subtests with names such as Similarities, Comprehension, Letter Matrices, Vocabulary, and that are specific to language and culture) made this list. The tests that are related to the flexibility of meanings, and word manipulation, also require abstract thinking abilities.

Shiu, Beaujean, Must, te Nienhuis, & Must (2013) has made an estimate the FE at the item level by using data from the Estonian National Intelligence Test (NIT) wherein it was found that approximately one-third of the subtest's items should be treated as a variant for FE comparisons. Non-invariant items were most frequently found in the Information and Analogies subtests (47–48% of items). In the Logical Selection (this subtest has different names, but in the NIT structure it is the third test of the A scale, so for our purposes we will call it the A3 subtest) 38% of items were non-invariant. This includes all of the last six items (from a total of 24). The content of the items of the A3 resembles the Similarities subtest of the original Binet and Stanford-Binet tests, and the Similarities subtest of the WAIS (Terman 1916; Wechsler 1955), the main aim of which was to evaluate abstract thinking ability. The item presentation order of the A3 is based on the order of difficulty (Whipple 1921). Shiu, Beaujean, Must, te Nienhuis, & Must (2013) eliminated 6 of the most difficult items due to their being non-invariant for FE comparisons. The reason for the lack of invariance of those items is unclear. It stands to reason that if 6 of the more difficult items were not included, then the real magnitude of the FE in this subtest would be suppressed.

Shiu, Beaujean, Must, te Nienhuis, & Must (2013) calculated the FE = 0.82 Hedges' g (Hedges 1981; based on the IRT methodology) for this subtest. Woodley of Menie et al. (2015) showed, that although word usage in modern societies is more extensive than in previous times, this may be attributed to the usage of a cognitively less demanding vocabulary. The lack of invariance of culturally loaded subtests may, however, not be inevitable. In the US, Beaujean & Sheng (2010) analyzed the invariance of the Vocabulary test of the General Social Survey 1972–2008, but did not detect a lack of invariance in the items.

3. Abstract thinking and the NIT

The NIT was developed in 1919–1920 and was based on the materials and methodologies of the Army Alfa and Beta Tests (Whipple 1921). The Army tests were developed in 1917–1918. The Army Alfa was a group, paper-pencil test consisting of eight subtests: oral directions, arithmetical reasoning, practical judgment, synonym-antonym, disarranged sentences, number series completion, analogies, information (Brigham 1923; Yoakum & Yerkes, 1920). Although the theoretical platform for developing the Army tests was the same as for the Stanford–Binet Scales (SBS), the Army tests did not include a subtest like the Similarities that was used in the SBS. However, an evaluation of abstract thinking ability is a central purpose of the SBS, and this includes its subtests such as Similarities, and Similarities Between Abstract Terms (Terman 1916). The main concern of those subtests is a comparison of objects based on their main features. It can be assumed that transforming the individually administered SBS directly into a formal paper and pencil format, which was necessary for the Army tests, was not an easy task for the test developers. This was the reason why the Army tests did not include a section analogous to the Similarities subtest that was used in the SBS. But a solution was soon found in the form of the development of the A3 subtest of the NIT. By content the items of the A3 (Haggerty et al., n.d.) resemble those of the Similarities subtest of the SBS, as they also require the separation of abstract qualities of an object (see the method part). In the process of adapting the test into the Estonian context, Tork (1940) re-named this subtest, and titled it "Concepts" as the items do not require formal logical operations, but rather require a comprehension of words and abstract thinking. The name of the original subtest makes use of the word *logical*, and by its nature logical analysis assumes abstractions. Special attention must be paid to the A3 subtest of the NIT in the context of the FE, as the subtest was developed in order to measure abstract thinking.