Personality and Individual Differences 50 (2011) 759–762

Contents lists available at ScienceDirect

Personality and Individual Differences

journal homepage: www.elsevier.com/locate/paid

Short Communication

Female Flynn effects: No sex differences in generational IQ gains

Jakob Pietschnig *, Martin Voracek, Anton K. Formann [1]

Department of Basic Psychological Research, School of Psychology, University of Vienna, Austria

ARTICLE INFO

Article history:
Received 23 July 2010
Received in revised form 6 December 2010
Accepted 14 December 2010
Available online 7 January 2011

Keywords:
Flynn effect
Verbal reasoning
Spatial ability
Mathematical reasoning
Sex differences

ABSTRACT

Generational changes of intelligence test performance in the general population (the Flynn effect) have been observed all over the world since the early 1940s. These changes are known to be country- and intelligence test domain-specific. To investigate whether such IQ gains are observable in three distinct domains of intelligence (verbal reasoning, spatial ability, mathematical reasoning) in German-speaking individuals, we examined a mixed-sex sample of 449 university students in a cross-sectional design. We assessed students' IQs on three original (standardized in 1970) and revised subscales (standardized in 2000) of a widely used German intelligence test battery, thus allowing investigation of test score changes over a time span of 30 years. Participants scored significantly higher on all subscales of the original test. Additionally, we observed higher performance of men than of women on all subscales, but only little evidence for sex differences regarding test score gains. In all, the Flynn effect appears to be progressive, robust, largely sex-independent, and intelligence domain-specific in respect to the magnitude of gains in German-speaking individuals.

1. Introduction

Generational IQ changes in the general population, termed the Flynn effect, have been subject to intensive investigation since the early 1980s (Flynn, 1984; Lynn, 1982). In the course of these examinations, a country-specific pattern emerged, indicating large gains in France, Japan, Israel, and the Netherlands, smaller ones in Great Britain, Ireland, New Zealand, and Australia, a stagnation in Norway and Sweden, and even a reversal in Denmark (e.g., Lynn, 2009; Teasdale & Owen, 2005).

Moreover, gains have been observed more frequently and more persistently in domains of fluid than crystallized intelligence in most countries and range typically from three to five IQ points per decade (Flynn, 1984, 2009; Lynn, 2009). However, an examination of crystallized IQ scores of German-speaking individuals showed robust gains of about two IQ points per decade (Voracek, 2006).

Usually, these effects are examined retrospectively using data that have been gathered incidentally. Typically, such data stem from examinations of IQ test performance of two or more large samples of men (foremost, military conscripts). Since identical test instruments are oftentimes used for a long time, it is possible to assess gains by comparing test performance of conscripts of different draft cohorts based on large representative samples displaying similar characteristics (e.g., Sundet, Borren, & Tambs, 2008). Although these designs usually comprise impressive numbers of participants, rather little is known about the Flynn effect in women. Since in most countries military service is mandatory for men only, only a single investigation of the Flynn effect in female military draftees in Israel is available (Flynn, 1998). Hence it remains unclear whether female Flynn effects display similar characteristics and accordingly should be investigated more thoroughly (Teasdale & Owen, 2005). In addition, these aforementioned designs have been criticized as problematic by some researchers, since the composition of samples may change over time and comparability of results is not warranted (Scott & Poncy, 1999).

To contribute to this important topic, we investigated gains over a time span of 30 years on three specific measures of intelligence covering a broad range of distinct abilities (verbal reasoning, spatial ability, mathematical reasoning). Up to date, these three intelligence domains have only been scarcely investigated in respect to the Flynn effect and thus should add novel evidence for this phenomenon, taking into account that intelligence test score changes across different countries appear to vary regarding their magnitude, direction, and domain (for a comprehensive review, see Flynn (2009)). In addition, studies addressing possible Flynn effects in spatial ability for Central Europe are unavailable.

By means of an established, but still underused, cross-sectional design (e.g., Satzger, Dragon, & Engel, 1996), we assessed intelligence test changes in a mixed-sex sample of university students. We hypothesized higher IQ scores of participants on all original subscales relative to the corresponding revised subscales. Taking previous findings into account (Flynn, 2009), we expected to observe strongest gains on subscales more closely associated with

* Corresponding author. Tel.: +43 1 4277 47842; fax: +43 1 4277 47849.
E-mail address: jakob.pietschnig@univie.ac.at (J. Pietschnig).
[1] Deceased.

0191-8869/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.paid.2010.12.019

fluid intelligence. Furthermore, we expected higher performance of men on all subscales, since men on average achieve slightly higher scores in these specific intelligence domains (Halpern, 2000, pp. 60, 70, 85, 211–212). By means of the present design we aimed to extend findings of the research literature on the Flynn effect to additional intelligence domains and to contribute to fill the gap in the literature regarding possible sex-specificity of Flynn effects.

2. Methods

2.1. Participants

To minimize influences of moderating variables, our goal was to attain a homogeneous sample in terms of sociodemographic characteristics. Accordingly, the total income-cohort of first-term psychology students at the University of Vienna was recruited, thus yielding a well-defined sample. Students received course credit for participation. Of total 449 participants (326f; $M = 21.6$ years, $SD = 3.8$, range: 18–49) 61.8% were of Austrian, 27.2% of German, and 11.0% of other nationality.

2.2. Materials

Three different subscales (Analogies, Cube Tasks, and Number Series) of the original and revised *Test of Intelligence Structure* (Amthauer, 1970; Amthauer, Brocke, Liepmann, & Beauducel, 2001), a well-known German intelligence test battery, were administered. These particular subscales were used because they cover a wide range of intelligence domains that have only been infrequently investigated regarding the Flynn effect and since test administration was deemed appropriate in terms of test economy. Moreover, this intelligence test battery is one of the most widely used and most carefully standardized test batteries in German. Although the item pool has been modified in the revised test in some cases, the two test forms are structurally equivalent as outlined below.

2.2.1. Analogies

The Analogies subscale is a measure of verbal reasoning. Each item presents three expressions of which the first two possess a certain relation. An analogous relation between the third expression and an additional one out of five possible alternatives has to be indicated (i.e., word 1:word 2 = word 3:?). All 20 items of the original and all 20 items of the revised test were administered (Cronbach α in above order: .55 and .55). Contentwise, there was no item overlap across original and revised subscales; however, all items were structurally identical.

2.2.2. Cube Tasks

By means of this subscale, spatial ability was assessed. Participants have to mentally rotate a stimulus cube and to indicate the matching rotated answer cube out of ten alternatives. The test material of the original and the revised subscale is identical; hence, the 20 items of this subscale were administered only once, but scored twice according to the 1970 and 2000 norms ($\alpha = .81$).

2.2.3. Number Series

In this subscale for mathematical reasoning, participants are presented a series of numbers following specific rules. To solve these tasks, participants have to identify the specific construction rule of the respective series and to provide its missing final number. All 20 items of the original and all 20 items of the revised subscale were administered ($\alpha = .91$ and .94). Two items were identical in original and revised subscales.

2.3. Procedure

Testing took place anonymously in a controlled classroom setting in groups of up to 40 participants in the presence of an experienced investigator and took about 50 min. To avoid influences of subscale order or standardization year, 12 different test booklets were prepared to allow counterbalancing and random assignment to test groups. Standardized test instructions were given verbally by the same test administrator for each group. Working time for each subscale was constrained to the respective suggested working time in the test manual (Analogies: 7 min each for the original and revised subscales; Cube Tasks: 9 min; Number Series: 10 min each for the original and revised subscales). Participants were asked to stop working after each subscale and to start the next tasks only when instructed to do so (i.e., when working time for the respective test had run out). After test completion, participants were thanked and debriefed via email.

3. Results

For analysis, participants' raw scores were standardized to IQ values according to the test-specific and age-specific norms of the respective subscale. Since original and revised test norms were established 30 years apart from each other, higher scores on the original test indicate gains on the respective subscale. This reasoning is appropriate, since the original and revised subscales are structurally equivalent and characterized by identical (Cube Tasks), overlapping (Number Series), or non-overlapping (Analogies) item content. This approach allowed treatment of performance on original and revised subscales as repeated observations in two points of time. Table 1 provides descriptive statistics for all subscales.

First, a multiple repeated-measures analysis of variance was calculated. Standardization year (original vs revised subscales) and subscale (Analogies vs Cube Tasks vs Number Series) were entered as within-subjects factors and sex as between-subjects factor. Standardization year reached significance, indicating differential performance of participants on original vs revised subscales (top of Table 2). Additionally, significant main effects for subscale and sex as well as interaction effects of sex with standardization year (differential change across the sexes) and subscale with standardization year (differential change across subscales) were observed, although variance proportions attributable to these interaction effects were negligible.

To further clarify these patterns, a series of univariate repeated-measures analyses of variance was calculated. Standardization year (original vs revised subscale) again served as the within-subject factor and sex as between-subjects factor. This procedure was

Table 1
Mean IQ and standard deviations of participants in subtests.

	Overall ($N = 449$)		Men ($N = 123$)		Women ($N = 326$)	
	Mean IQ	*SD*	Mean IQ	*SD*	Mean IQ	*SD*
Analogies						
1970	110.0	11.1	112.9	9.9	108.9	11.3
2000	108.2	12.2	109.7	11.4	107.7	12.5
Cube Tasks						
1970	98.9	18.2	102.6	18.0	97.6	18.2
2000	95.1	15.3	98.5	15.3	93.8	15.1
Number Series						
1970	115.3	19.5	121.0	17.6	113.1	19.8
2000	106.7	20.4	109.2	19.1	105.8	20.8
Overall						
1970	108.1	11.5	112.2	10.4	106.6	11.6
2000	103.3	10.9	105.8	10.0	102.4	11.1

Table 2
Analyses of variance.

	F	*df*	*p*	η_p^2
Multiple repeated-measures analysis of variance				
Sex	17.8	1, 447	<.001	.04
Standardization year	119.4	1, 447	<.001	.21
Subscale	105.0	2, 446	<.001	.32
Sex × standardization year	5.4	1, 447	.021	.01
Sex × subscale	1.3	2, 446	.267	.01
Standardization year × subscale	13.5	2, 446	<.001	.06
Sex × standardization year × subscale	2.0	2, 446	.131	.01
Univariate repeated-measures analyses of variance				
Analogies				
Standardization year	14.5	1, 447	<.001	.03
Sex	7.8	1, 447	.005	.02
Standardization year × sex	2.8	1, 447	.097	<.01
Cube Tasks				
Standardization year	517.2	1, 447	<.001	.54
Sex	7.6	1, 447	.006	.02
Standardization year × sex	0.8	1, 447	.358	<.01
Number Series				
Standardization year	55.2	1, 447	<.001	.11
Sex	11.6	1, 447	.001	.03
Standardization year × sex	2.9	1, 447	.090	<.01

Note: Subscale corresponds to Analogies, Number Series, and Cube Tasks.

carried out separately for Analogies, Cube Tasks, and Number Series (bottom of Table 2).

IQ scores on subscales as well as the overall score were significantly higher for the original than for the revised test (all *ps* < .001). Gains were strongest for Number Series (2.9 IQ points per decade), moderate for Cube Tasks (1.3 per decade), and smallest for Analogies (0.6 per decade). Average overall IQ gains per decade were 2.1 points for men (ranging from 1.1 to 4.0 across subscales), 1.4 for women (0.4 to 2.4), and 1.6 for the total sample (0.6 to 2.9).

There were significant sex differences on all subscales (bottom of Table 2; Fig. 1), revealing higher test scores among men (all *ps* < .01). However, no interaction of sex and standardization year was observed on either subscale (all *ps* > .05). Subgroup analyses for native Austrian vs German participants yielded virtually identical results (omitted for brevity).

4. Discussion

In the present study, we demonstrate a robust Flynn effect for three intelligence domains in a German-speaking sample. The effect is differentiated (i.e., different domains show different gains), ranging from about half a point per decade for verbal reasoning, over more than one point for spatial ability, to almost three points for mathematical reasoning. These findings present several points of interest, as elaborated in the following.

First, participants invariably scored higher on all three original subscales than on the revised ones. This indicates that the Flynn effect is remarkably robust regarding different domains of intelligence. These results conform to previous findings from Central Europe (Voracek, 2006).

Second, although gains were observed across all tested domains, strength of gains differed. Smallest gains were found

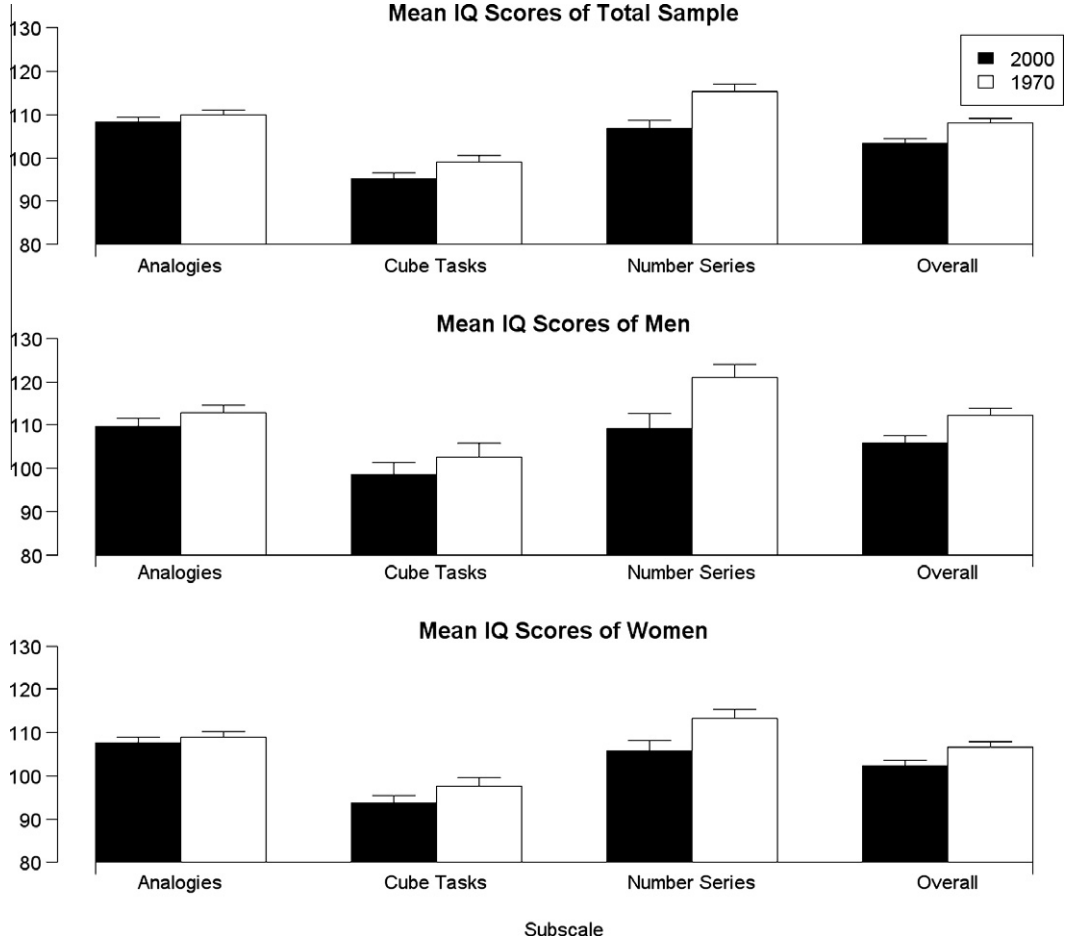

Fig. 1. Mean IQ of men and women on the original and revised subscales.

for verbal reasoning and turned out to be smaller than gains observed in an Estonian sample by Must, Must, and Raudik (2003) of about two points per decade. Although in our investigation the reliability of this subscale was low, this emphasizes robustness of our results: due to the nature of our statistical models, in presence of lower power of tests, more pronounced mean differences are needed to yield significant results.

To the authors' knowledge, there has been no study yet addressing the Flynn effect in spatial ability in Central Europe. Recent evidence of a study examining Flynn effects in measures of spatial ability in a Swedish sample (Rönnlund & Nilsson, 2008) indicated gains in this domain. Unfortunately, the authors of this study reported only overall changes in cognitive task performance (1.5 points per decade) and did not report gains for this specific domain. However, gains of the present study are comparable in magnitude (1.3 points per decade) to overall gains as described by Rönnlund and Nilsson (2008).

Gains in the Number Series subscale in our sample were strongest (2.9 points per decade) and of comparable magnitude as usually demonstrated on measures of fluid intelligence. These results conform to findings from most European countries and the United States, where typically gains of about three points per decade are observed (Flynn, 2009; Lynn, 2009).

Overall intelligence test score gains in our sample were moderate, amounting to 1.6 IQ points per decade. Gains were observable on all subscales, thus suggesting a pervasive, but still domain-specific, Flynn effect. As hypothesized, they were strongest on subscales closely related to fluid intelligence, displaying moderate gains in Analogies, stronger ones in Cube Tasks, and strongest in Number Series. These results emphasize the need for the use of more diverse and comprehensive test measures for differing domains of intelligence in future studies. Of note, amounts of explained variance differed to a rather large extent between subscales, indicating decreasing effects of identical item content of subscales regarding unobserved variance.

Third, sex differences on task performance in all subscales were observed, yielding significantly higher mean IQ scores of men (ranging from 2.0 to 7.9 points). Of note, largest sex differences were observed for mathematical reasoning and not for spatial ability. However, all employed subscales are measures of intelligence domains for which it is well-known that observed test performance of men typically is higher (Halpern, 2000).

It should be noted that our design did not allow more detailed assessment of changes over time (e.g., curvilinearity) but assumes linearity, resulting in estimates of average changes per decade. However, this was accepted to allow investigation of the little addressed subject of sex-specificity of the Flynn effect (Teasdale & Owen, 2005).

Moreover, this cross-sectional design is a highly valuable approach for reliable identification of Flynn effects, as it allows target-orientated examination of distinct domains of intelligence (Scott & Poncy, 1999). Additionally, assessment of within-subject differences on all three subscales completely rules out threats to validity due to sampling error. As the present subscales represent excellently validated and psychometrically well-examined test instruments, threats to validity of our findings due to different item difficulties in original and revised subscales can be largely ruled out.

To summarize, our results provide evidence for a progressive Flynn effect in German-speaking individuals. This effect appears to be robust, largely sex-independent, and differentiated across domains of intelligence.

References

Amthauer, R. (1970). *Intelligence structure test*. Göttingen, Germany: Hogrefe.

Amthauer, R., Brocke, B., Liepmann, D., & Beauducel, A. (2001). *Intelligence Structure Test I-S-T 2000 R*. Göttingen, Germany: Hogrefe.

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

Flynn, J. R. (1998). Israeli military IQ tests: Gender differences small; IQ gains large. *Journal of Biosocial Science, 30*, 541–553.

Flynn, J. R. (2009). *What is intelligence? Beyond the Flynn effect* (expanded paperback ed.). Cambridge, UK: Cambridge University Press.

Halpern, D. (2000). *Sex differences in cognitive abilities* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

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

Lynn, R. (2009). Fluid intelligence but not vocabulary has increased in Britain, 1979–2008. *Intelligence, 37*, 249–255.

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

Rönnlund, M., & Nilsson, L.-G. (2008). The magnitude, generality, and determinants of Flynn effects on forms of declarative memory and visuospatial ability: Time-sequential analyses of data from a Swedish cohort study. *Intelligence, 36*, 192–209.

Satzger, W., Dragon, E., & Engel, R. R. (1996). The equivalence of the German version of the Wechsler Adult Intelligence Scale-Revised (HAWIE-R) and the original German version (HAWIE). *Diagnostica, 43*, 119–138.

Scott, R., & Poncy, B. (1999). The "Flynn Effect": How applicable is it to longitudinal IQ assessment of American university students? *Mankind Quarterly, 39*, 385–397.

Sundet, J. M., Borren, I., & Tambs, K. (2008). The Flynn effect is partly caused by changing fertility patterns. *Intelligence, 36*, 183–191.

Teasdale, T. W., & Owen, D. R. (2005). A long-term rise and recent decline in intelligence test performance: The Flynn effect in reverse. *Personality and Individual Differences, 39*, 837–843.

Voracek, M. (2006). Phlogiston, fluid intelligence, and the Lynn–Flynn effect. *Behavioral and Brain Sciences, 29*, 142–143.