No negative Flynn effect in France:

Why variations of intelligence should not be assessed using tests based on cultural knowledge

Corentin Gonthier

Université Rennes 2

Jacques Grégoire

Université Catholique de Louvain

Maud Besançon

Université Rennes 2

Corentin Gonthier, Université de Rennes, LP3C EA 1285, 35000 Rennes, France

Jacques Grégoire, Université Catholique de Louvain, 1348 Louvain-la-Neuve, Belgique

Maud Besançon, Université de Rennes, LP3C EA 1285, 35000 Rennes, France

Correspondence concerning this article should be addressed to Corentin Gonthier, Laboratoire LP3C, Campus Villejean, Place du Recteur Henri Le Moal, CS 24307, 35043 Rennes Cedex, France. E-mail: corentin.gonthier@univ-rennes2.fr

Word count: 10.541 words excluding title page, abstract, figures and tables

No negative Flynn effect in France: Why variations of intelligence should not be assessed using tests based on cultural knowledge

1. Introduction

In 2015, Dutton and Lynn published a study claiming that intelligence scores in France are decreasing. This study was based on data collected by the publisher of Wechsler's adult intelligence scale (WAIS; Wechsler, 2000, 2011). In the process of developing the WAIS-IV in 2009, the publisher collected a sample of 79 subjects who performed both WAIS-III and WAIS-IV, in order to ensure that the two versions had convergent validity. Dutton and Lynn (2015) compared performance averages reported in the test manual for this validity sample, and observed that these subjects had higher standardized scores on the WAIS-IV (in reference to its 2009 normative sample) than on the WAIS-III (in reference to its 1999 normative sample). Using the 79 subjects performing both versions as a common reference point, they interpreted this finding as evidence that the 1999 WAIS-III normative sample had higher average ability than the 2009 WAIS-IV normative sample (for details on this reasoning, see Flynn, 1998b). Dutton and Lynn concluded that intelligence had declined from 1999 to 2009 - hence a reversal of the Flynn effect (Flynn, 1984; Rundquist, 1936). A commentary of Woodley of Menie and Dunkel (2015) insisted that this decline was likely due to biological causes.

Unusually for a psychological study, Dutton and Lynn's work received much attention from mainstream French media. Translating a few headlines found in the French press can help measure the extent of the moral panic created by this particular article: *The IQ of the French in freefall*; *Alert: the IQ of Asians skyrockets, ours decreases*; *Vertiginous decrease of IQ: researchers raise the alarm*; and the most publicized, a full-length documentary on a major TV channel: *Tomorrow: we will all be morons*. The decline of intelligence in France has become a common topic of conversation for laypeople, and a mandatory question from psychology undergraduates during introductory courses on intelligence, all of this based on the Dutton and Lynn (2015) study. Explanations for the decline proposed as "feasible" by Dutton and Lynn, including a flood of low-IQ immigrants, have also been hotly debated.

Given the societal impact of this work, further scrutiny seems warranted. Is the decline of intelligence reported by Dutton and Lynn (2015) the whole story? There has been a recent trend of articles illustrating intelligence decreases; a systematic literature review reported such findings in eight samples spanning seven different countries (see Dutton, van der Linden, & Lynn, 2016). In some cases, these studies included very large sample sizes, up to all conscripts of a country in a given year (Dutton & Lynn, 2013; Shayer et al., 2007; Sundet et al., 2004; Teasdale & Owen, 2004). On the other hand, recent large meta-analyses have substantially disagreed on this matter: one meta-analysis found that the Flynn effect has slowed but not halted (Pietschnig & Voracek, 2015; including 4 million subjects in 31 countries), and another concluded that the Flynn effect continues at the same rate (Trahan et al., 2014; including 14.000 subjects in 285 studies). Both meta-analyses listed a small number of studies reporting negative Flynn effects, which suggests that observed intelligence decreases can reflect fluctuations around a stable or increasing average (see Pietschnig & Voracek, 2015, Figure 2; Trahan et al., 2014, Figure 2), due either to chance or to methodological bias.

1.2. Three Outstanding Issues and the Role of Cultural Knowledge

1.2.1. Methodological Issues

There are a number of methodological reasons to reserve judgment about the data reported by Dutton and Lynn (2015) in particular. The first reason is the sample size of 79 subjects, small enough for a study on the Flynn effect that replication is warranted: random patterns of variation are a strong possibility with such a limited sample. A related question about sampling is the lack of information in the test manual about the way the sample was collected: because this was a very secondary part of the process of developing the WAIS-IV, the publisher did not provide much detail on sample composition, making it unclear whether this was a representative sample of the French population (for a similar point criticizing the use of the same procedure by Flynn in 1998, see Zhu & Tulsky, 1999).

The second methodological reason to reserve judgment is the fact that the authors could not perform any significance test: the publisher reported only averages and standard deviations for the whole sample in the test manual, but the design was within-subjects (in which case inferential statistics require access to the raw data to compute the variance of difference scores). For the same reason, the effect sizes reported in Dutton and Lynn (2015) could only be computed as if the design had been between-subjects. Given the limited sample size, this absence of statistical tests makes it unclear whether the difference between WAIS-III and WAIS-IV was in fact significantly greater than chance, and if so, which subtests of the WAIS were affected.

1.2.2. Differences between Subtests: Fluid versus Crystallized Intelligence

Besides methodological issues, Dutton and Lynn's (2015) interpretation considers all subtests of the WAIS as interchangeable measures of intelligence. However, it is well known that different indices of the WAIS actually reflect different narrow dimensions of ability (for factor analyses, see Benson et al., 2010; Flanagan, 2000; Flanagan et al., 2013; Golay & Lecerf, 2011; Keith et al., 2006; Lecerf et al., 2012; Ward et al., 2012). In the terms of the Cattell-Horn-Carroll (CHC) taxonomy of intelligence (McGrew, 2009), the Matrix Reasoning subtest of the Perceptual Reasoning Index (PRI) reflects *Gf* (fluid intelligence: inductive and deductive reasoning abilities), and the Picture Completion and Block Design subtests of the

PRI reflect *Gv* (visual processing). Vocabulary, Similarities, Information and Comprehension, the four subtests of the Verbal Comprehension Index (VCI), all reflect *Gc* (crystallized intelligence: acquired knowledge within a specific culture, and application of this knowledge). Subtests of the Processing Speed Index and Working Memory Index reflect *Gs* (processing speed) and *Gsm* (short-term memory), respectively. Lastly, the Arithmetic subtest appears to reflect a mix of *Gf*, *Gc*, *Gsm* and *Gq* (quantitative reasoning).

In other words, all subtests of the WAIS do not equally assess intelligence: subtests of the VCI (which require e.g. defining words or remembering trivia) are largely reflections of culture-based declarative knowledge, whereas subtests of the PRI depend more on inductive reasoning and visuo-spatial ability. This is not to say that these indices represent orthogonal dimensions of ability: all subtests of the WAIS load on the general factor of intelligence to an extent (Canivez & Watkins, 2013), and a *Gf* subtest such as Matrix Reasoning depends in part on procedural knowledge related to manipulation of abstract materials, which also varies across cultures (see e.g. Van de Vijver, 2016); conversely, *Gc* subtests such as Vocabulary require subjects to elaborate a response in a way that goes beyond simple retrieval of knowledge learned by heart. Still, the weight of declarative knowledge is substantially greater in VCI subtests reflecting *Gc*, as illustrated by their lesser decline with age (Baxendale, 2011, Grégoire, 1993; Kaufman et al., 1989, Ryan et al., 2000) and their higher correlation with education and socio-economic levels (Dori & Chelune, 2004; Heaton et al., 2003; Holdnack et al., 2013), than all other subtests that place little emphasis on declarative knowledge.

Critically, careful examination of the data reported by Dutton and Lynn (Table 3) suggests that the difference between WAIS-III and WAIS-IV was mainly driven by subtests depending on *Gc* (Vocabulary, Comprehension, Information), suggesting that it did not reflect a general decline of intelligence. Although the lack of statistical tests in their study made this point uncertain, such a result would severely limit interpretation of the findings in terms of a "negative Flynn effect". The first reason for this is societal: when laypersons hear talk of a decline of intelligence, they do not tend to think about a "decline of culture-based knowledge". On the contrary, decline of intelligence is clearly portrayed in the media as a decline of logical reasoning, which should primarily appear on subtests reflecting *Gf*. The second reason hinges on the fact that the generational gains in intelligence labeled *Flynn effect* primarily occur on *Gf*, much more than *Gc* (Pietschnig & Voracek, 2015; for results with the Wechsler scales, see Grégoire et al., 2016). If the finding of a decline in scores was restricted to subtests with a *Gc* component, the logical conclusion would be that it reflects the operation of a different mechanism than the classic Flynn effect - presumably a mechanism related to acquired declarative knowledge, whose role is greater on these subtests. The classic Flynn effect may also be caused, in part, by environment-driven changes in knowledge, but if this is the case it presumably has to do more with procedural knowledge related to manipulation of abstract test material than with the kind of declarative knowledge required by *Gc* subtests (e.g. Flynn, 1998a).

1.2.3. Emerging Item Bias and Cultural Changes in the WAIS

A final, related issue is that interpreting the difference between WAIS-III and WAIS-IV as a difference of ability rests on the assumption that the WAIS-III and WAIS-IV measure ability equally well in a recent sample. As stressed by Beaujean and Zheng (2014): "*this process is predicated on the belief that different editions of the same instrument measure the same construct(s), the same way. [...] These between-edition mean comparisons are akin to comparing average temperatures at two different geographic locations with thermometers that use different scales*". (For further discussion, see Kaufman, 2010; Nugent, 2006; Weiss et al., 2015; Zhu & Tulsky, 1999). Another useful analogy is that of estimating individual differences in height (intelligence) by measuring the length of shadows (test scores): this method can yield useful estimates of individual differences in height at a given time, but comparisons between shadows collected at different seasons will be biased (Flynn, 1998a; Jensen, 1994).

In this view, the difference in average standardized scores between WAIS-III and WAIS-IV could reflect a difference in their measurement properties for the normative samples and for the validity sample completing both versions, instead of a difference in the average intelligence of their respective normative samples (see also Rodgers, 1998). In line with this idea, intelligence tests - including the WAIS - do not demonstrate measurement invariance over time: in particular, the baseline difficulty of a given subtest can change over time, even when controlling for ability (Wicherts et al., 2004). In other words, even the same version of the same test can vary in the way it reflects the underlying construct when performed by different cohorts at different dates; measurement bias can appear over time. Wicherts and colleagues (2004) found that this measurement bias often affected subtests with a substantial weight of culture-based declarative knowledge; in their study with the WAIS, the two subtests with substantial measurement bias across cohorts were Similarities and Comprehension, both subtests of the VCI.

This idea of measurement bias appearing over time due to changes of culture-based declarative knowledge is best understood at the item level. For instance, asking subjects to define the same word will yield less and less correct answers over time if that particular word falls into disuse, independently of the subjects' ability. On a similar note, Wicherts (2007) gives the example of Dutch subjects failing to answer a WAIS item about the Kremlin by responding that it is "a small, cute, furry creature", a confusion caused by the release of Steven Spielberg's film *The Gremlins*. The performance of Dutch subjects may have decreased, but it would be absurd to conclude that their intelligence has decreased... unless one is willing to accept the converse example given by Wicherts: that a higher average performance to the WAIS item asking to define the word "terminate" following the release of the movie *Terminator*, would mean that the release of *Terminator* increased average intelligence.

These examples constitute an instance of item bias, which happens when different subjects with the same level of ability have a different probability of answering the item correctly, due to an extraneous variable irrelevant to the construct being measured. Such bias leads to differences of scores without a corresponding difference of ability. In the context of differences between groups or cohorts, item bias is usually framed as differential item functioning (DIF): a group of subjects can demonstrate lower performance, not because these subjects have lower ability, but because the items themselves are biased to be intrinsically more difficult for this particular group due to an extraneous variable (see e.g. Ackerman, 1992; Martinková et al., 2017; Zumbo, 2007).

In line with this idea, prior research has shown that some items in intelligence tests tend to become significantly easier or harder over time (item drifts; e.g. Brand et al., 1989). The framework of item response theory (IRT) is particularly suited to address the question of this item bias, as it allows for deconfounding item characteristics and participants' ability (see e.g. Beaujean & Osterlind, 2008). Studies using this approach have shown that item parameters do vary across cohorts, with drifts of item difficulty for some items and loss of discriminating power for others, especially for tests involving culture-based knowledge such as vocabulary and mathematics (Beaujean & Osterlind, 2008; Pietschnig et al., 2013). When considering total performance, two studies found that observed intelligence scores changed over several decades, while latent ability as estimated from IRT remained near-constant (Beaujean & Osterlind, 2008; Beaujean & Sheng, 2010), confirming both that item properties can change without corresponding changes of ability, and that this can create the illusion of ability changes over time.

If the WAIS-III items have become outdated, it is easy to understand why subjects performing the test now would perform lower than they would on the more recent WAIS-IV, and comparatively lower than the normative sample who completed the WAIS-III when it was first designed, creating the illusion of a score decline. Critically, all subtests composing the WAIS are not equally vulnerable to the possibility of item properties changing over time. Subtests primarily reflecting *Gf*, such as Matrix Reasoning, or *Gv*, such as Block Design, are based on abstract materials which should be relatively timeless; the same is true for subtests reflecting *Gs* and *Gsm*. On the other hand, subtests primarily reflecting *Gc* intrinsically depend on culture-based declarative knowledge, and culture as assessed by these tests changes over time (e.g. Pietschnig et al., 2013), making them especially likely to develop DIF (see Beaujean & Osterlind, 2008; Pietschnig et al., 2013; Wicherts et al., 2004).

The possibility of item bias emerging over time, due to cultural changes, thus provides a plausible mechanism to explain why the difference between WAIS-III and WAIS-IV reported by Dutton and Lynn (2015) could have been primarily driven by subtests with a large *Gc* loading: Vocabulary (define words, whose prevalence in the language can evolve), Similarities (find the common feature of two verbally presented concepts, which can be more or less familiar in a given cultural context), Information (answer questions about general knowledge, whose representation in school curricula and in the media can change), Comprehension (find justifications for social and cultural rules, which can be more or less stressed in daily life or even fall out of use), and Arithmetic (mentally perform arithmetic problems, based on operations which can become less familiar as school curricula and daily life activities change).

Consistent with this idea, the content of all five subtests was significantly modified between the WAIS-III and WAIS-IV: the French publisher changed 11 of the 18 Comprehension items, 29 of the 33 Vocabulary items, 21 of the 28 Information items, and 12 of the 19 Similarities items. More importantly, in-depth comparison of the items indicates substantial differences in item content between the two versions. For example, the WAIS-III Information subtest includes 12 items asking about the identity of famous people (five of which lived and died in the twentieth century), whereas the WAIS-IV version includes only 4 items about famous people, mostly at lower difficulties. In the WAIS-III Comprehension subtest, 7 of the most difficult items concern civic education and economy, whereas the WAIS-IV version includes only 1 such item; conversely, the WAIS-IV version includes 5 items concerning ecology and development aid, topics which were absent from the WAIS-III. Other changes that could create difficulty in recent samples are more subtle: for example, 8 out of the 20 items in the WAIS-III Arithmetic subtest require subjects to calculate prices expressed in *francs*, the former French currency which was replaced by *euros* in 2002. These changes would not be sufficient, in and of themselves, to explain the difference reported by Dutton and Lynn[1], but they could inflate the difference between WAIS-III and WAIS-IV, and the fact that such extensive changes were deemed necessary hints that items of the older WAIS-III may be less appropriate for a recent sample.

1.3. Research Questions

The purpose of the current study was to re-examine the possibility of a negative Flynn effect in France, using the same approach as Dutton and Lynn (2015), but controlling for the

[1] The validity sample of 79 subjects performed lower on the WAIS-III than on the WAIS-IV, in reference to their respective normative samples. In theory, this use of normed scores accounts for procedural differences: based on these results, it would be valid to conclude that the WAIS-III normative sample had higher average ability than the WAIS-IV normative sample, even if the two had performed entirely different tests - *provided that both tests were equally unbiased, indifferent indicators of intelligence*. On the other hand, if the WAIS-III displays measurement bias against recent samples, the only possible result assuming constant ability is the one reported by Dutton and Lynn: the recent validity sample will necessarily perform lower on the WAIS-III than on the WAIS-IV relative to their normative samples. This would be true even with identical items for the WAIS-III and WAIS-IV (on average, the 2009 normative sample would perform less well than the 1999 normative sample on the same items designed in 1999 due to bias), but using more up-to-date items on the WAIS-IV will increase the difference even further (the performance of the validity sample will be both decreased on the WAIS-III due to bias, and comparatively enhanced on the updated items of the WAIS-IV).

possibility that the reported decline of intelligence reflected a spurious difference, driven by item bias emerging over time on subtests with a large component of culture-based declarative knowledge.

This question can be summarized in the context of a hierarchical model of performance on intelligence tests (for an example, see Wicherts et al., 2004; Wicherts, 2007). Performance can be represented as a hierarchical model with four levels: on the top is latent general intelligence, *g*; on the second level are latent broad dimensions of intelligence such as *Gf*, *Gc*, *Gq, Gs* and *Gsm*; on the third level is observed performance on the various subtests that serve to estimate latent dimensions; and on the fourth level is observed performance on the specific items that constitute the subtests. The question asked here comes down to asking at what level exactly resides the difference reported by Dutton and Lynn (2015). The Flynn effect is supposed to be a latent increase at the first level of *g*, especially prevalent for the second level factor of *Gf*; and this increase is expected not to be caused by measurement bias at the third or fourth levels (see Flynn, 2009a). By contrast, the results of Dutton and Lynn suggest both that there is no decline for the first level of *g*, and a limited decline for the second level but only for *Gc*; and that this decline could be caused, not by an actual change of latent ability, but by measurement bias at the fourth level representing items.

Testing this possibility required answering the three major issues detailed above: 1) ensuring that the sample analyzed by Dutton and Lynn was appropriate and adding significance tests, so as to 2) confirm that the decline essentially appeared for those subtests primarily reflecting *Gc*, and 3) determine whether the decline on these subtests was caused by differential item functioning - item bias associated with higher difficulty for subjects in a recent sample, with the same level of ability.

Study 1 re-analyzed the same dataset as Dutton and Lynn (2015), using the raw data to which they did not have access. To this end, we obtained permission to use the original data from the test publisher. We complemented the data with inferential tests and appropriate effect sizes, so as to determine whether the difference between WAIS-III and WAIS-IV was significant for all subtests, or whether it was driven by subtests with a large *Gc* loading (Vocabulary, Similarities, Information, Comprehension, Arithmetic), as the data seemed to suggest. We also verified the composition of the sample (the distribution of age and IQ in particular) to ensure that there were no issues that could affect the results.

Study 2 aimed to replicate and extend the results of Study 1 by collecting a new dataset ($N = 79$). This was partly intended as a replication effort to compensate for the small sample size of Study 1; we conducted analyses both on the new sample, and on the combination of Study 1 and Study 2 samples (for a total $N = 155$) to ensure stability of the results. Another purpose was to collect item-level data, which were not recorded for the Study 1 dataset: this made it possible to determine whether differences between versions could be driven by DIF on certain items of the WAIS-III, reflecting systematic bias against a recent sample. The French publisher also authorized access to the item-level data for the WAIS-III 1999 normative sample ($N = 1104$); item difficulty parameters for the subtests demonstrating a decline were estimated, and compared with the new sample collected in 2019 for Study 2. This analysis made it possible to test whether a given item had a different probability of being solved correctly in the two samples, for a subject with the same level of ability.

As a complementary step, we also investigated the conclusion of Woodley of Menie & Dunkel (2015) that the decline of performance in France could be biological given correlations between performance decline and indices of biological load. We first re-examined the data used by the authors to estimate the biological load of WAIS subtests and re-analyzed their results, and we then considered the issue from another angle, by searching for a correlation between performance decline and cultural load (see Kan et al., 2013).

2. Study 1

2.1 Method

2.1.1. The dataset

The French publisher authorized access and use of the raw data for the validity sample collected to assess convergent validity between the WAIS-III and WAIS-IV. As described in the test manual and in Dutton and Lynn (2015), the dataset comprised 79 subjects (mean age = 44.53 years, *SD* = 13.71, range = 30 to 63 years). Each subject completed both the WAIS-III and WAIS-IV in counterbalanced order (40 subjects performed the WAIS-III first and 39 performed the WAIS-IV first). Dutton and Lynn (2015) raised concerns regarding the interval between the two sessions, but this information was not recorded. Performance on each subtest was computed in reference to the respective normative samples of each version (*N* = 1104 for the WAIS-III, collected in 1999, and *N* = 876 for the WAIS-IV, collected in 2009).

Additional information not available in the manual could also be retrieved from the test publisher, and is provided here for archival purposes. This validity sample was collected in the process of developing the WAIS-IV, by psychologists hired by the publisher to collect this validity data on a paid-per-protocol basis. All psychologists received specific training from the publisher prior to data collection; each psychologist sent back protocols to the publisher to ensure that they complied with data collection instructions and that the test was scored correctly. Each subject completed the WAIS-III and WAIS-IV with the same psychologist. Subjects who completed the WAIS-IV first (*n* = 39) were included in the normative sample for the WAIS-IV, and there was additional information available for them (sample composition should be similar for the other half of the sample). These 39 subjects were assessed by 22 different psychologists. Most psychologists assessed a single subject, others up to 6 subjects. Subjects were recruited using the method of quotas, with quotas on age, gender, and socio-economic level. This subsample included 19 men and 20 women; socio-economic level was assessed based on the categories of the French national institute of statistics (*INSEE*), and appeared to match the composition of the general population (to be precise, this subsample included 2 farmers, 3 artisans, 5 executives and other intellectual professions, 8 workers of intermediate level, 7 employees, 7 laborers, and 7 students or unemployed persons; retired persons are counted as per their former profession). Geographic regions in this subsample were somewhat unbalanced, with 28 of the 39 subjects coming from southern France.

2.1.2. Data analysis

For each subtest and each index, we first compared the WAIS-III and WAIS-IV performance using a within-subjects *t*-test. The corresponding within-subjects effect sizes were computed using Cohen's *dz* (computed as the average of differences between WAIS-III and WAIS-IV, divided by the standard deviation of these differences; Cohen, 1988; see also Lakens, 2013). We report both uncorrected *p*-values, and significance after correction for multiple comparisons; correction was performed using the Benjamini-Hochberg method (false discovery rate or FDR: fixing the probability of making at least one type I error over all comparisons at 5%; this method is less severe and more powerful than the more common Bonferroni correction when there are multiple significant effects; see Benjamini & Hochberg, 1995). A second series of analyses tested the difference between WAIS-III and WAIS-IV using mixed-design ANOVAs, including all possible interactions with test order and age group (see Results section 2.2.2 for this variable) as between-subjects variables; the corresponding effect sizes were computed as partial eta squared. Contrasts analyses were used to decompose significant interactions.

2.2 Results

2.2.1. Complementing the Data Analysis with Significance Tests

For consistency, we first re-analyzed the data as they were presented in the test manual and as they were interpreted by Dutton and Lynn (2015). We just corrected two errors prior to analysis. The first error was a misreporting for the Arithmetic subtest in the WAIS-IV manual (WAIS-IV performance reported as $M = 10.1$, $SD = 3.0$ instead of $M = 10.63$, $SD = 3.24$). The second error was a mistranslation on the part of Dutton and Lynn (in their Table 4, line 5; the numbers they report as "Perceptual Organization Index" actually refer to a different comparison involving verbal IQ, which is not meaningful here). There were missing values for the Digit Symbol Coding subtest, yielding unequal sample sizes across tests.

Descriptive statistics and the corresponding inferential tests are reported in Table 1. The results confirmed that the full-scale IQ (FSIQ) was significantly lower on the WAIS-III than on the WAIS-IV ($p < .001$), compatible with a decrease of intelligence between the WAIS-III and WAIS-IV normative samples. Contrary to this hypothesis, however, the difference was mostly driven by *Gc* subtests: Vocabulary, Information and Comprehension subtests, along with the corresponding VCI, all showed medium-to-large (Cohen, 1988) performance decreases (*dz* between .54 and .62). A significant difference also appeared for Block Design, an index of *Gv*, but with a small-to-medium effect size (*dz* = .33). The Matrix Reasoning subtest, used as the best available index of *Gf* (Marshalek et al., 1983; Carpenter et al., 1990), had an even smaller effect size and showed a non-significant difference ($p > .10$) after correction for multiple comparisons. Subtests and indices reflecting working memory and processing speed were unaffected.

A complementary analysis using linear regression indicated that the predicted difference of FSIQ between versions for a subject with a null difference in the VCI was small and not significantly different from zero, $b = -0.12$, $t = -0.07$, $p = .947$, confirming that *Gc* subtests accounted for most or all of the difference in FSIQ. In sum, there was a significant difference between WAIS-III and WAIS-IV in FSIQ, but it was mostly attributable to a difference in subtests with a large *Gc* loading: incompatible with a reversal of the Flynn effect classically observed to a greater extent on *Gf* subtests and also incompatible with a general decline of intelligence, but compatible with measurement bias on subtests involving culture-based declarative knowledge.

Table 1

Study 1 data as analyzed by Dutton & Lynn (2015), corrected and with significance tests

Measure	*N*	WAIS-III scores (SD)	WAIS-IV scores (SD)	*t*	*p*	*dz*	*Corrected sig. (BH)*
Vocabulary	79	8.77 (2.68)	9.94 (2.90)	5.41	<.001	0.61	***
Similarities	79	9.94 (2.89)	10.09 (3.01)	0.55	.585	0.06	
Information	79	8.70 (3.18)	9.82 (3.02)	5.17	<.001	0.58	***
Comprehension	79	8.73 (2.99)	9.85 (2.81)	4.89	<.001	0.55	***
Picture completion	79	9.92 (3.76)	10.34 (3.07)	0.92	.360	0.10	
Block design	79	9.91 (3.47)	10.57 (3.14)	2.90	.005	0.33	*
Matrix reasoning	79	9.59 (3.41)	10.14 (3.02)	1.91	.060	0.22	
Arithmetic	79	10.01 (2.72)	10.63 (3.24)	2.26	.026	0.25	°
Digit span	79	10.24 (3.09)	10.22 (2.49)	-0.10	.923	0.01	
Letter-number sequencing	79	10.13 (3.14)	10.23 (2.91)	0.31	.757	0.03	
Digit symbol coding	73	9.68 (3.42)	9.63 (2.88)	-0.20	.844	0.02	
Symbol search	79	10.53 (4.23)	10.33 (3.75)	-0.42	.673	0.05	
Verbal comprehension index	79	95.14 (13.82)	99.87 (14.92)	5.48	<.001	0.62	***
Perceptual reasoning index	79	98.84 (17.63)	102.00 (16.28)	2.42	.018	0.27	*
Working memory index	79	100.67 (14.82)	102.34 (13.56)	1.68	.098	0.19	
Processing speed index	74	100.65 (18.18)	101.01 (16.59)	0.19	.847	0.02	
Full scale IQ	74	98.72 (14.16)	102.39 (14.95)	4.08	<.001	0.47	***

Note. t refers to a within-subjects Student's *t*-test; *p* is the uncorrected *p*-value; *dz* is a within-subjects version of Cohen's *d* effect size; *Corrected sig. (BH)* indicates significance after Benjamini-Hochberg correction for multiple comparisons (****p*<.001, ***p*<.01, **p*<.05, °*p*<.10), applied separately to subtests and summary indices.

2.2.2. Re-analysis of the Data, Taking into Account Sample Composition

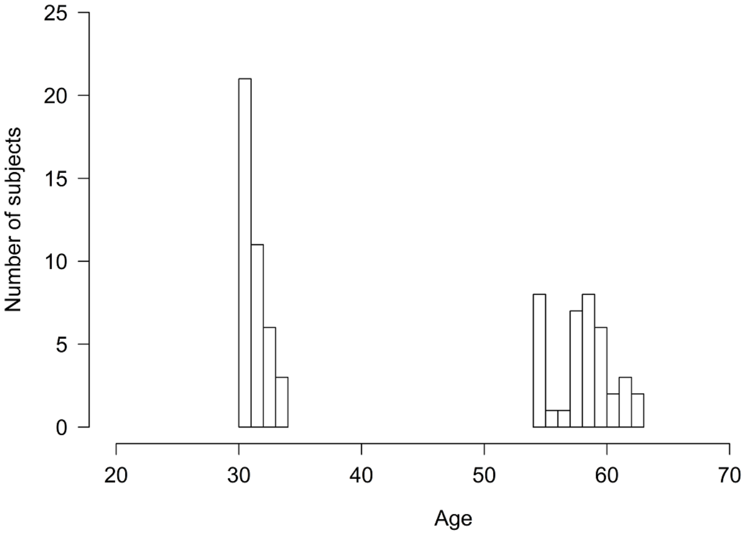
Apart from the additional details on sample composition summarized in the Methods (section 2.1.1), two aspects of sample composition invited further scrutiny. The first was the age of subjects in the sample. The descriptive statistics for age, as stated in the test manual and in Dutton and Lynn (2015), were $M = 44.53$ years, $SD = 13.71$ years, range = 30 to 63 years. This is technically correct, but overlooks the very peculiar composition of the sample, represented in Figure 1. As is immediately visible, the data were actually collected in two discrete age groups: 30-to-34 and 55-to-63 year-olds. (Note that the counterbalancing with test order was performed correctly: 20 younger and 20 older subjects performed the WAIS-III first, whereas 21 younger and 18 older adults performed the WAIS-IV first.)

From the point of view of the publisher, recruiting a younger and an older group was a reasonable choice: this validity sample was collected only to ensure that the WAIS-III correlated with the WAIS-IV, and the publisher wished to ensure that this was the case for both younger and older adults; instead of collecting a few subjects in all age groups, they collected more subjects in two extreme age groups and checked that the correlation between versions was high in the two groups. Because this validity sample was never intended to be analyzed for other purposes, there was no particular incentive to make it representative. Thus, the sample in this dataset was definitely not representative of the general population as a whole, but there was no reason to expect that it should be the case (see also Zhu & Tulsky, 1999).

We also examined the distribution of IQ scores in the sample, which revealed that out of 79 subjects, five reached the threshold for intellectual disability (FSIQ less than 70; the lowest IQ in the sample was 57, about three standard deviations below average) on both the WAIS-III and WAIS-IV. This represented a larger prevalence of intellectual disability than expected in the general population (6.3% of the sample instead of an expected 2.5%), and yielded an unbalanced range of ability given that there were no gifted subjects at the other end of the scale (the highest FSIQ in this sample was 123). Again, including these subjects may have made sense for the publisher as a way to ensure that both versions of the test yielded similar conclusions about intellectual disability, but it would make more sense to exclude them from a dataset testing for the Flynn effect - the absence of intellectual disability should reasonably have been an exclusion criterion for such a study. Indeed, a major issue is that these five subjects were outliers on both the Matrix Reasoning subtest and the Block Design subtest, with a standard score of 1 (the minimum possible, reflecting complete failure to meet task requirements), but only in the WAIS-III. This discrepancy can happen when different versions of the same test have different discriminating power for subjects with low performance, due for instance to different stopping criteria; in the present case, it largely contributed to the observed difference between WAIS-III and WAIS-IV for these two subtests (Table 1), potentially biasing the results.

Figure 1

Actual distribution of ages in the Study 1 sample analyzed by Dutton and Lynn (2015)



Given the peculiar composition of the sample, we performed a new series of analyses after excluding the five subjects with intellectual disability. The results ($N = 74$) are displayed in Table 2 and in Figure 2a. The major changes with this corrected sample were that the difference between WAIS-III and WAIS-IV for the Arithmetic subtest became significant even after correction for multiple comparisons, whereas the difference between WAIS-III and WAIS-IV on Matrix Reasoning (marginally significant in the whole sample, without correction for multiple comparison) completely disappeared, uncorrected $p = .214$. At the level of indices, only the VCI and FSIQ showed a statistically significant difference between versions at the .05 level; both survived correction for multiple comparisons. These results confirmed the prior conclusion that the observed difference between WAIS-III and WAIS-IV was almost exclusively driven by *Gc* subtests, whereas no difference appeared for most other subtests, including the *Gf* subtest of Matrix Reasoning, as well as working memory and processing speed subtests.

Table 2

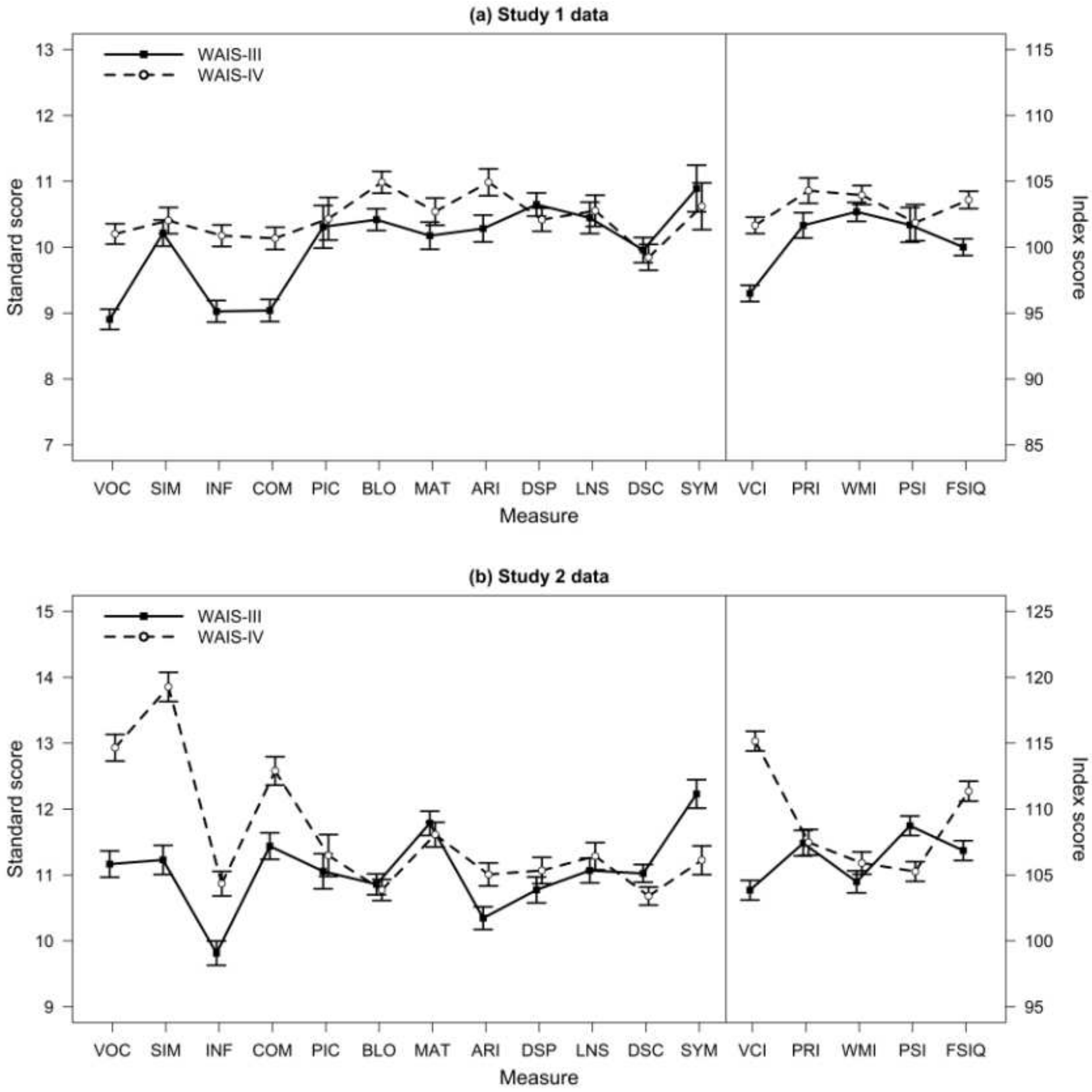
Study 1 data after exclusion of five subjects with intellectual disability

Measure	*N*	WAIS-III scores (SD)	WAIS-IV scores (SD)	*t*	*dz*	*p*	*Corrected sig. (BH)*
Vocabulary	74	8.91 (2.63)	10.20 (2.79)	5.94	0.69	<.001	***
Similarities	74	10.22 (2.65)	10.41 (2.82)	0.67	0.08	.504	
Information	74	9.03 (3.00)	10.18 (2.78)	4.98	0.58	<.001	***
Comprehension	74	9.04 (2.78)	10.14 (2.62)	4.60	0.53	<.001	***
Picture completion	74	10.31 (3.54)	10.43 (2.98)	0.27	0.03	.791	
Block design	74	10.42 (2.87)	10.99 (2.73)	2.43	0.28	.017	*
Matrix reasoning	74	10.18 (2.65)	10.54 (2.66)	1.25	0.15	.214	
Arithmetic	74	10.28 (2.59)	10.99 (3.02)	2.44	0.28	.017	*
Digit span	74	10.65 (2.64)	10.42 (2.37)	-0.93	0.11	.358	
Letter-number sequencing	74	10.45 (2.96)	10.55 (2.55)	0.32	0.04	.749	
Digit symbol coding	70	9.99 (3.13)	9.87 (2.69)	-0.41	0.05	.686	
Symbol search	74	10.89 (4.06)	10.62 (3.68)	-0.54	0.06	.592	
Verbal comprehension index	74	96.50 (12.91)	101.66 (13.59)	5.88	0.68	<.001	***
Perceptual reasoning index	74	101.66 (14.13)	104.30 (13.82)	1.91	0.22	.060	
Working memory index	74	102.69 (12.94)	103.96 (12.23)	1.25	0.14	.217	
Processing speed index	71	102.03 (17.02)	102.20 (15.87)	0.09	0.01	.931	
Full scale IQ	71	100.31 (12.05)	103.97 (13.04)	3.92	0.47	<.001	***

Note. t refers to a within-subjects Student's *t*-test; *p* is the uncorrected *p*-value; *dz* is a within-subjects version of Cohen's *d* effect size; *Corrected sig. (BH)* indicates significance after Benjamini-Hochberg correction for multiple comparisons (****p*<.001, ***p*<.01, **p*<.05, °*p*<.10), applied separately to subtests and summary indices.

Figure 2

Difference between scores normed on the WAIS-III and WAIS-IV in Study 1 and Study 2



Note. VOC = Vocabulary, ARI = Arithmetic, SIM = Similarities, DSP = Digit Span, COM = Comprehension, INF = Information, LNS = Letter-Number Sequencing, MAT = Matrix Reasoning, SYM = Symbol Search, DSC = Digit Symbol Coding, PIC = Picture Completion, BLO = Block Design, VCI = Verbal Comprehension Index, PRI = Perceptual Reasoning Index, WMI = Working Memory Index, PSI = Processing Speed Index, FSIQ = Full Scale IQ. Errors bars represent within-subjects standard errors of the mean (Morey, 2015).

We also found it of interest to determine whether the difference between WAIS-III and WAIS-IV varied as a function of age group (younger group vs. older group), and/or as a function of test order (WAIS-III first vs. WAIS-IV first). A series of ANOVAs indicated that for most measures, taking into account test order and age group did not substantively change Dutton and Lynn's conclusions. Most interactions involving age and test order were non-significant (*p*s > .05). For Digit Symbol Coding and Picture Completion, test version interacted with test order ($p = .026$ and $p < .001$ respectively), but these were simply crossover interactions indicating a test-retest effect, wherein the test version performed first yielded lower scores than the test version performed second.

The only substantial differences appeared for the Block Design and the Arithmetic subtests. Block Design demonstrated both a two-way interaction between test version and test order, $F(1, 70) = 5.60$, $p = .021$, $\eta^2_p = .07$, and a three-way interaction between test version, age group, and test order, $F(1, 70) = 4.34$, $p = .040$, $\eta^2_p = .06$. Arithmetic demonstrated both a two-way interaction between test version and age group, $F(1, 70) = 4.97$, $p = .029$, $\eta^2_p = .07$, and a marginal two-way interaction between test version and test order, $F(1, 70) = 3.30$, $p = .073$, $\eta^2_p = .05$. A contrast analysis indicated the same pattern for both subtests: only older adults who completed the WAIS-III first performed significantly higher on the WAIS-IV ($p < .001$); there was no difference between WAIS-III and WAIS-IV for younger adults in either order condition, or for older adults who performed the WAIS-IV first (all other *p*s > .43 for Block Design and all other *p*s > .24 for Arithmetic).

2.3 Discussion

The first finding of Study 1 was that the sample composition was poorly suited to address the specific question of Flynn effects, due to a design with two discrete age groups (Figure 1), and the presence of five individuals with intellectual disability (6.3% of the sample) who completely failed some subtests of the WAIS-III. The second finding, after re-analyzing the data, was that the difference between WAIS-III and WAIS-IV was statistically significant with a medium effect size, but was almost entirely driven by three of the four subtests of the VCI - in other words, three subtests primarily depending on *Gc*. No difference appeared for the sole *Gf* subtest, Matrix Reasoning, or for working memory or processing speed.

After exclusion of subjects with intellectual disability, significant differences also remained for the Arithmetic subtest, which also includes a significant *Gc* loading, and for Block Design, a *Gv* subtest, but these differences were more unstable and only appeared in the subgroup of older adults who performed the WAIS-III first. It is unclear why differences would appear only in this subgroup. A likely possibility is random variation due to the small sample size (there were only 20 subjects in this particular condition); another possible explanation is that the WAIS-III provided training that makes it easier to deal with items of the WAIS-IV, and that older adults benefit particularly from this training. In any case, this finding questioned the meaning and the stability of the difference between WAIS-III and WAIS-IV for these two particular subtests, and invited replication in Study 2.

In sum, of the five subtests demonstrating a difference between WAIS-III and WAIS-IV, all except Block Design involved a significant contribution of declarative knowledge, and the difference for the latter only existed in one of four subgroups. In fact, all subtests involving cultural component showed a difference except for Similarities. This is clearly incompatible with the existence of a negative Flynn effect reflecting a general decline of intelligence. Based on these data, it would be more accurate to say that this sample provided evidence for a recent decline of performance on tests depending on cultural knowledge and designed at the end of the 1990s.

3. Study 2

3.1 Method

3.1.1 Participants

Study 2 was designed as a replication of Study 1, meaning data collection was planned for the same sample size. A total of 81 subjects, recruited by word of mouth in the community, participated in the study. As in Study 1, subjects were included using the method of quotas, with quotas on gender, age, and socio-economic level. The sample included 42 women and 39 men, mean age = 38.44 years, *SD* = 11.94, range = 20 to 60 years (with a continuous distribution). Socio-economic groups were representative of the general population (1 farmer, 3 artisans, 17 executives and other intellectual professions, 19 workers of intermediate level, 20 employees, 15 laborers, and 6 students or unemployed persons; retired persons counted as per their former profession). All subjects were native French speakers, and none had a history of major neurologic or psychiatric disorder.

3.1.2 Procedure

Subjects completed both the WAIS-III and WAIS-IV (Wechsler 2000, 2011) in counterbalanced order (WAIS-III first: *n* = 42; WAIS-IV first: *n* = 39). The median test-retest interval was 35 days (*M* = 35.92 days, *SD* = 14.92 days, range = 14 to 70 days), similar for the two orders (median = 35 days for both). Data collection was performed by 24 young psychologists as a part of their final graduate training (we reasoned that their level of expertise with the WAIS-IV would be similar to that of psychologists paid to collect the data during WAIS-IV development in Study 1); all had received extensive training with Wechsler scales. Each experimenter collected 2 to 5 protocols, and each subject completed the WAIS-III and WAIS-IV with the same experimenter.

3.1.3 Data Analysis

As in Study 1, the first series of analyses testing the difference between WAIS-III and WAIS-IV used within-subjects *t*-tests, with *dz* effect sizes and FDR correction for multiple comparisons. A second series of analyses combined the Study 1 and Study 2 datasets, so as to increase power to detect differences between WAIS-III and WAIS-IV; these analyses were performed using mixed-design ANOVAs, with test version (WAIS-III vs. WAIS-IV) as a within-subjects variable, and sample (Study 1 vs. Study 2) included as a between-subjects variable to account for average differences of performance, also controlling for the two-way interaction between test version and sample. Taking advantage of the increased sample size for this combined sample, we also computed Bayes Factors representing the likelihood of the alternative hypothesis of a difference between WAIS-III and WAIS-IV over the null (with uniform priors). In particular, this allowed us to quantify evidence in favor of the null hypothesis, for subtests other than the five subtests reflecting *Gc*.

The second purpose of Study 2 was to investigate DIF between the 2019 sample collected here ($N = 81$) and the reference 1999 normative sample for the WAIS-III ($N = 1104$). Ability estimates were computed from WAIS-III subtests that did not show a difference between samples. The analysis then used IRT to model predicted performance on a given item as a function of ability, yielding an item characteristic curve; this was done separately for the two samples. Lastly, we compared the probability of solving each item correctly for the same level of ability across the two samples, using Raju's signed area method (estimating the area between the two item characteristic curves). This analysis was performed for the five subtests that demonstrated a significant decrease between the WAIS-III and WAIS-IV[2].

Subjects in the 2019 sample performed somewhat above the 1999 average, and there was a greater range of scores in the 1999 sample, especially for lower ability levels. This could confound comparison between the two samples (with item characteristic curves being estimated as a function of ability and some regions of ability existing only in one of the two samples). To control for this possibility, a preliminary step for the DIF analysis was to equate the range of ability in the two samples by removing subjects of the 1999 sample who performed lower or higher than all subjects of the 2019 sample. The final sample size for the 1999 sample was *n* = 767.

Item parameter estimation was performed using Stocking's method A (Stocking, 1988; see also Birnbaum, 1968; Ban et al., 2001). Instead of jointly estimating item parameters and latent subject ability, as is commonly the case in IRT, this method uses a known estimate of subject ability to calibrate item parameters. We used as an estimate of subject ability an index of general intelligence, computed as the average of standardized performance on all subtests of the WAIS-III (with standardization performed conjointly over the two samples), excluding the five subtests being tested for DIF (to avoid the possibility of measurement bias influencing the ability estimate). This general ability estimate correlated between .41 and .60 with the five biased subtests. This approach had several advantages in the present situation:

[2] As is common with tests of DIF, alternative methods yielded different results and the list of items identified with DIF varied, but this did not change the overall picture much. For example, a test of DIF using logistic ordinal regression, with ability estimates computed more classically using IRT based on the subtest being tested for DIF (using package *lordif*; Choi, 2016) found significant bias for the five subtests: vocabulary (12 biased items), similarities (10 biased items), information (10 biased items), comprehension (6 biased items), and arithmetic (4 biased items). When aggregated at the test level, this led to bias against the 2019 sample for all subtests. Another analysis using the Mantel-Haenszel method, with all items recoded as binary and with ability computed as total score on a subtest, also found significant bias for vocabulary (11 biased items), similarities (3 biased items), information (8 biased items), comprehension (1 biased item), and arithmetic (3 biased items), mostly against the 2019 sample. Both methods are probably less accurate than the one used here in this context.

Stocking's method A works well with limited sample sizes, as was the case for the 2019 sample; it removes the necessity of picking anchor items to link the ability scales of the two groups on the same metric; and it estimates general ability based on all available information, instead of only the items being tested for DIF. The latter propriety meant that the ability estimate was not confounded with measurement bias in the items being tested, and that the same ability estimate was used to test DIF in the five subtests.

Estimation of item parameters using this method was performed with the package *irtplay* 1.4.1 (Lim, 2020) for *R* (R Core Team, 2020), separately for the reference 1999 normative sample and the current 2019 sample. Due to the low sample size in the 2019 sample, we only examined difficulty parameters, with the discrimination parameter fixed to 1: the one-parameter logistic model (1-PL; see Birnbaum, 1968; Rasch, 1960) was used for items scored 0-1, and the partial credit model (PCM with a slope fixed to 1; Masters, 1982) was used for items scored 0-1-2. Items which were not answered by a subject (due to the WAIS rule of discontinuing a test after the subject fails several successive items) were scored 0 prior to parameter estimation.

After estimating item difficulty parameters separately for the two samples, a decision statistic summarizing the extent of DIF was computed using Raju's signed area method, which consists in integrating the area between the two item characteristic curves (see Raju, 1988; note that this is equivalent to computing the difference between samples in the level of ability required to obtain a predicted 0.5 score when items are scored 0 or 1). This approach was chosen over alternatives, such as the Mantel-Haenszel method or logistic regression (e.g. Zumbo, 2007), for two reasons: 1) it appeared to be the most straightforward solution to analyze subtests including a mix of items scored 0-1 and items scored 0-1-2, and 2) contrary to most other methods, it could be naturally extended to quantify differential test functioning (bias at the test level) by integrating the item characteristic curves for all items.

Inferential statistics were obtained using the item parameter replication method (Oshima et al., 2006), with the correction proposed by Cervantes (2012, 2017a; see also Clark & LaHuis, 2012). For each item, difficulty parameters were randomly generated for the two samples under the null hypothesis (drawn from a normal distribution, with the same mean equal to the value of the difficulty parameter in the reference 1999 sample, and with variance and covariance equal to their respective values in the 1999 sample and in the 2019 sample) using package *mvtnorm* for R (Genz et al., 2019). This operation was repeated 5.000 times for each item to generate a vector of signed areas between item characteristic curves under the null hypothesis; the *p*-value of the signed area for each item was then computed as the corresponding quantile in this vector (two-tailed).

As a final step for the analysis of differential functioning, item characteristic curves for all items of a subtest were combined to create a test characteristic curve, reflecting the predicted score as a function of ability at the subtest level. This made it possible to assess the aggregate effect of DIF over all items. The extent of differential functioning at the test level was summarized by computing the difference of predicted total score between the two samples, for an average ability. We also converted this difference of predicted performance into a difference of Wechsler standard scores ($M = 10$, $SD = 3$), so as to place it on the same scale as Tables 1-3 to get a sense of how it compared to the size of the difference between WAIS-III and WAIS-IV observed in the 2019 validity sample. Note that this conversion is only a rough approximation, as the conversion from raw scores to standard scores is not linear (it uses a normalized scale, which means multiple raw scores yield the same standard score) and varies as a function of age and level of ability (we used the norms for 35 to 44 year-olds on the WAIS-III as a point of reference corresponding to the mean age of our sample).

3.2 Results

3.2.1 Effect of test version: Replication of Study 1

Differences between performance on the WAIS-III and WAIS-IV were generally in line with the results of Study 1 (see Table 3 and Figure 2b). As in Study 1, there was a significant difference of FSIQ between the two versions; in Study 2, this difference was driven exclusively by the five subtests reflecting *Gc*: there were significant differences for Vocabulary, Similarities, Information, Comprehension, the corresponding VCI, and the Arithmetic subtest. Four of these replicated from Study 1; the only change in Study 2 was for the Similarities subtest, which had not shown a significant difference in Study 1. Subtests from the PRI did not show any difference, all *p*s > .500; in particular, the effect on Block Design observed in Study 1 did not replicate, uncorrected $p = .704$, and performance on the *Gf* subtest of Matrix Reasoning did not differ between WAIS-III and WAIS-IV, uncorrected $p = .515$. One unexpected finding is that subjects performed significantly higher on Symbol Search in the WAIS-III version than on the WAIS-IV, suggesting an *increase* in processing speed between the WAIS-III and WAIS-IV samples, even more incompatible with a negative Flynn effect.

Table 3

Study 2 data

Measure	*N*	WAIS-III scores (SD)	WAIS-IV scores (SD)	*t*	d_z	*p*	*Corrected sig. (BH)*
Vocabulary	81	11.16 (2.51)	12.94 (3.42)	6.24	0.69	<.001	***
Similarities	81	11.24 (2.73)	13.84 (2.99)	8.40	0.93	<.001	***
Information	81	9.80 (2.78)	10.88 (3.30)	4.04	0.45	<.001	***
Comprehension	81	11.39 (2.76)	12.63 (2.91)	4.01	0.45	<.001	***
Picture completion	81	10.98 (3.12)	11.39 (2.91)	0.45	0.05	.655	
Block design	81	10.87 (2.86)	10.77 (2.55)	-0.38	0.04	.704	
Matrix reasoning	81	11.78 (2.31)	11.62 (2.61)	-0.65	0.07	.515	
Arithmetic	81	10.32 (2.95)	11.04 (2.89)	2.71	0.30	.008	*
Digit span	81	10.77 (2.80)	11.07 (3.05)	1.05	0.12	.296	
Letter-number sequencing	81	11.12 (3.23)	11.23 (2.94)	0.78	0.09	.438	
Digit symbol coding	81	11.02 (2.70)	10.68 (2.88)	-1.79	0.20	.078	
Symbol search	81	12.22 (2.93)	11.23 (3.15)	-3.28	0.36	.002	**
Verbal comprehension index	81	103.84 (12.40)	115.16 (15.44)	10.72	1.19	<.001	***
Perceptual reasoning index	81	107.37 (12.96)	107.53 (12.75)	0.04	0.00	.965	
Working memory index	81	104.41 (14.95)	105.98 (15.00)	1.20	0.13	.235	
Processing speed index	81	108.71 (13.40)	105.30 (15.13)	-3.30	0.37	.001	**
Full scale IQ	81	106.80 (11.87)	111.41 (13.55)	4.22	0.47	<.001	***

Note. t refers to a within-subjects Student's *t*-test; *p* is the uncorrected *p*-value; *dz* is a within-subjects version of Cohen's *d* effect size; *Corrected sig. (BH)* indicates significance after Benjamini-Hochberg correction for multiple comparisons (****p*<.001, ***p*<.01, **p*<.05, °*p*<.10), applied separately to subtests and summary indices.

3.2.2 Effect of test version: Combining the Study 1 and Study 2 Datasets

The next analysis combined the datasets from Study 1 and Study 2, both to ensure the stability of the results, and to ensure that the fact differences between WAIS-III and WAIS-IV were restricted to certain subtests was not due to insufficient power. The results of ANOVAs for this complementary analysis are summarized in Table 4; they were very similar to Study 1 and Study 2 considered separately.

As in Study 2, the five subtests with significant cultural influence (Vocabulary, Similarities, Information, Comprehension, and Arithmetic) all elicited significantly lower performance in the WAIS-III version than in the WAIS-IV version. There were no differences for the other subtests, apart from a slightly higher performance in the WAIS-IV version of Symbol Search, driven by the Study 2 sample. The same pattern emerged for summary indices, with a significant difference between versions only for the VCI reflecting *Gc*. There was also a marginal difference between versions for the Working Memory Index, but it was driven exclusively by the Arithmetic subtest which also involves declarative knowledge.

Bayesian analyses confirmed that there was very strong evidence in favor of differences between the WAIS-III and WAIS-IV for the same five subtests depending on *Gc*, and only for these subtests. The Bayes factor was not informative for Symbol Search, thus diverging from the frequentist results: this did not support the existence of a large difference between versions for this subtest. Critically, there was substantial evidence in favor of the null hypothesis for the three subtests assessing *Gf* and *Gv* - Matrix Reasoning, Picture Completion and Block Design - and for the corresponding PRI, as well as for other subtests reflecting *Gsm* and *Gs*. In sum, the combined datasets of Study 1 and Study 2 were firmly incompatible with the possibility of a negative Flynn effect reflecting a general decrease of intelligence or a larger decrease for *Gf* subtests, and instead confirmed that performance decreased specifically for the subtests involving declarative knowledge.

Table 4

ANOVA table for the combination of Study 1 and Study 2 data

Measure	*N*	*F*	η^2_p	*p*	*Corrected sig. (BH)*	BF_{10}
Vocabulary	155	71.41	.32	<.001	***	2.69 x $10e^{11}$
Similarities	155	44.32	.22	<.001	***	3.20 x $10e^{6}$
Information	155	39.38	.20	<.001	***	3.02 x $10e^{6}$
Comprehension	155	35.46	.19	<.001	***	608010
Picture completion	155	0.25	.00	.620		0.14
Block design	155	2.19	.01	.141		0.28
Matrix reasoning	155	0.24	.00	.625		0.13
Arithmetic	155	13.19	.08	<.001	***	57.28
Digit span	155	0.03	.00	.860		0.12
Letter-number sequencing	155	0.35	.00	.553		0.15
Digit symbol coding	151	1.89	.01	.171		0.32
Symbol search	155	4.92	.03	.028	*	1.38
Verbal comprehension index	155	141.21	.48	<.001	***	9.60 x $10e^{18}$
Perceptual reasoning index	155	1.91	.01	.169		0.26
Working memory index	155	2.89	.02	.090		0.52
Processing speed index	152	2.41	.02	.122		0.47
Full scale IQ	152	32.22	.18	<.001	***	207855

Note. The table reports the main effect of test version (WAIS-III vs. WAIS-IV) in a series of ANOVAs controlling for sample (study 1 vs. study 2). *Corrected sig. (BH)* indicates significance after Benjamini-Hochberg correction for multiple comparisons (***p<.001, **p<.01, *p<.05, °p<.10), applied separately to subtests and summary indices ; BF_{10} indicates the Bayes Factor corresponding to the likelihood of the alternative hypothesis (WAIS-III ≠ WAIS-IV) versus the null hypothesis.

3.2.3 Analysis of DIF

The final series of analyses investigated whether lower performance on the five subtests of the WAIS-III relying on cultural knowledge was attributable to DIF for a recent sample - in other words, to items functioning differently when compared to the reference 1999 sample, rather than a lower ability in the 2019 sample. The analysis of DIF was performed on subtests where the average difference between WAIS-III and WAIS-IV was significant in Study 2 (except for Symbol Search, which does not involve discrete items). Tests of DIF for all items are summarized in Table 5, and item characteristic curves for the items with the most bias against the 2019 sample are presented in Figure 3a for illustration.

For the Vocabulary subtest, 10 out of 33 items were significantly biased against the 2019 sample, meaning they required comparatively higher intellectual ability to be solved for subjects in the 2019 sample compared to subjects in the 1999 sample. Of these 10 items, two required ability over one standard deviation higher for subjects in the recent sample, for the same probability of being solved correctly as in 1999. Two more items were biased against the 2019 sample at the trend level. On the other hand, 4 out of 33 items were significantly easier for the 2019 sample than for the 1999 sample, compatible with the idea of evolving patterns of word use in the language (e.g. Brand, 1989).

For the Similarities subtest, 3 out of 19 items were significantly biased against the 2019 sample. The most bias was obtained for Item 15, which required ability almost one standard deviation higher in the 2019 sample; this makes particular sense because this item involves customs, which have practically disappeared in France in recent years due to the opening of borders within the European Union, and the French word for the other concept required by this item is now explicitly labeled as "dated" in many dictionaries. Item 17 also required ability half a standard deviation higher in the 2019 sample; this item involves the process for obtaining rubber, which may be less obvious to subjects now than it was over 20 years ago. On the other hand, 2 out of 19 items were easier in the recent sample.

For the Information subtest, 11 out of 28 items were significantly more difficult in the 2019 sample, and none were significantly easier. Almost all biased items concerned the identification of famous people; this was the case for six out of the seven items that required ability at least one standard deviation higher in the 2019 sample, with four of these items involving famous writers.

For the Comprehension subtest, item parameter estimation was impossible for 12 out of 18 items due to too few failures in the 2019 sample. Out of the remaining six items for which estimation was possible, two items were significantly more difficult in the 2019 sample, and a third item was marginally more difficult. All three items concerned civic education, dealing with the topics of parliaments and criminal courts.

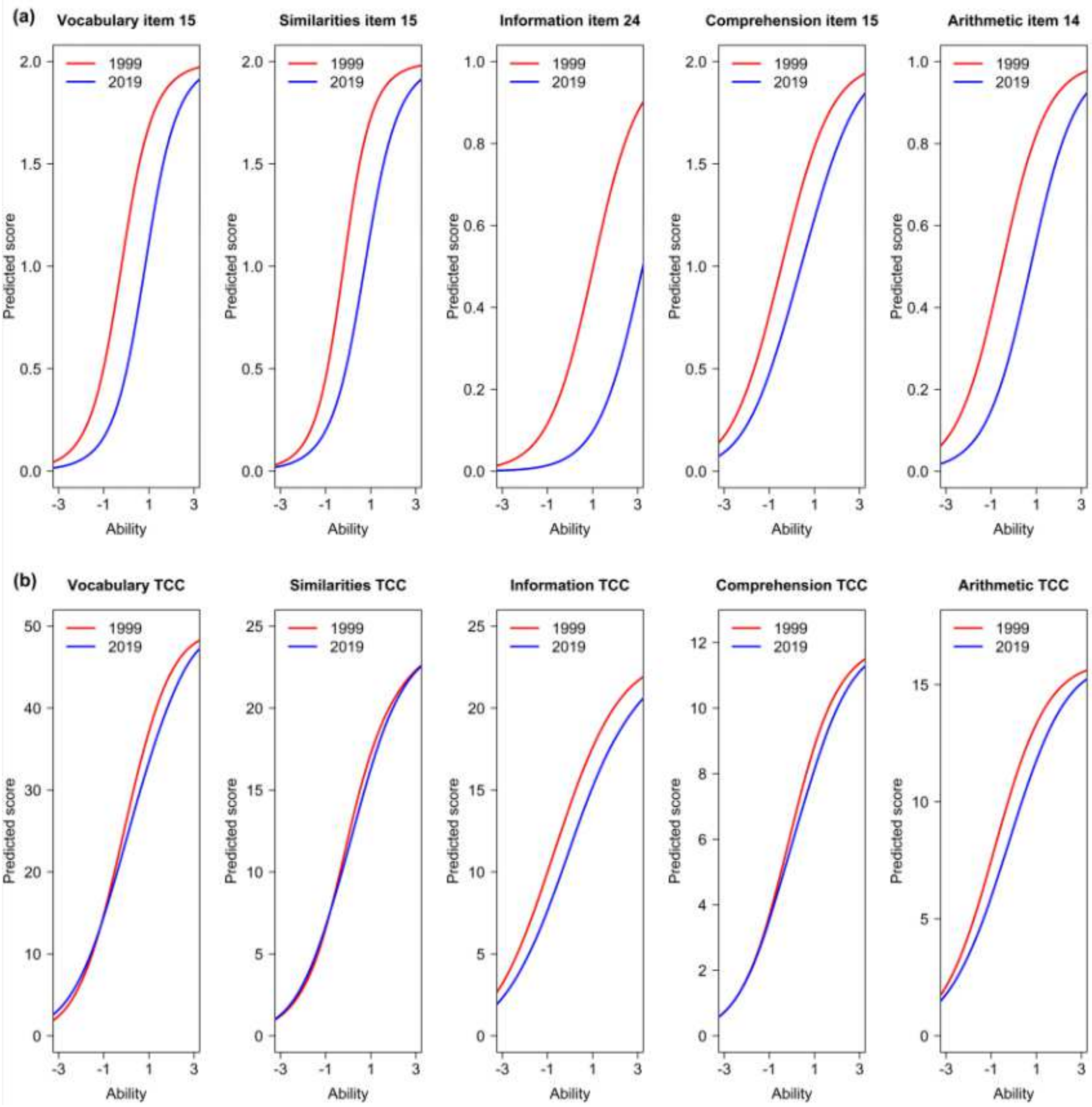
Lastly, for the Arithmetic subtest, 7 out of 20 items were significantly more difficult for the 2019 sample, and none were significantly easier. Two items required ability more than one standard deviation higher in the 2019 sample, and both required mental division; out of the other five biased items, four also required mental division, either explicitly or in the context of computing a proportion.

The five test characteristic curves for these five subtests are represented in Figure 3b. Characteristic curves for Vocabulary, Information and Arithmetic showed substantial bias against the 2019 sample; bias was somewhat smaller for the Similarities subtest, and for the Comprehension subtest though this may be due to the very low number of items available for parameter estimation. The data corresponding to these test characteristic curves are summarized in Table 6 (for example, obtaining half the points on the Vocabulary subtest required ability 0.21 standard deviations higher in the 2019 sample than in the 1999 sample; with an average ability, the predicted raw score was 2.30 points lower in the 2019 sample, corresponding to a standard score roughly 1 point lower). As reflected in Table 6, for all five subtests, differential test functioning elicited scores 0.5 to 1 point lower in terms of Wechsler standard scores (about 0.17 to 0.33 standard deviations lower) for the 2019 sample than for the 1999 sample, with an equal level of ability.

Comparing this result with the size of the difference between WAIS-III and WAIS-IV suggested that differential test functioning explained much or all of the score difference interpreted by Dutton and Lynn (2015) as a negative Flynn effect. This was the case for all subtests except perhaps for Similarities, where differential test functioning between the 1999 and 2019 sample accounted for about one third the size of the observed difference between WAIS-III and WAIS-IV. We reiterate that these are not precise estimates, but a rough approximation based on extrapolating WAIS-III norms, computed just to get a sense of relative effect sizes. It is also not entirely meaningful to interpret the difference between WAIS-III and WAIS-IV based on differential test functioning for the WAIS-III, given that the WAIS-IV is based on very different materials.

Figure 3

Item characteristic curves and test characteristic curves for the five subtests showing differential functioning



Note. (a) Item characteristic curves for the item with the largest bias against the 2019 sample in each subtest, and (b) Test characteristic curve (TCC) for each subtest. The x-axis represents standardized ability as computed from all subtests except the five represented here. For ease of reading, the curves are plotted in *red* for the *reference* 1999 group and in blue for the 2019 group.

Table 5

Differential item functioning for the five subtests associated with score declines

Item	Vocabulary		Similarities		Information		Comprehension		Arithmetic	
	Diff.	*p*	Diff.	*p*	Diff.	*p*	Diff.	*p*	Diff.	*p*
Item 01	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC
Item 02	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC
Item 03	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC
Item 04	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC
Item 05	NC	NC	NC	NC	NC	NC	NC	NC	-0.085	.830
Item 06	NC	NC	NC	NC	0.501	.325	NC	NC	0.429	.148
Item 07	NC	NC	NC	NC	-0.061	.886	NC	NC	-0.432	.477
Item 08	**0.494**	**.033**	**-0.469**	**.037**	**0.868**	**.003**	NC	NC	0.291	.342
Item 09	-0.276	.195	0.016	.928	-0.018	.960	NC	NC	0.202	.488
Item 10	-0.478	.317	0.023	.908	-0.230	.647	NC	NC	-0.046	.894
Item 11	-0.001	.997	-0.187	.321	0.303	.328	NC	NC	0.340	.213
Item 12	NC	NC	-0.171	.377	**1.036**	**.001**	NC	NC	0.435	.132
Item 13	0.231	.174	**0.473**	**.001**	**1.199**	**<.001**	-0.209	.259	**0.727**	**.006**
Item 14	**-0.624**	**.008**	0.242	.126	**0.786**	**.004**	0.292	.097	**1.276**	**<.001**
Item 15	**1.051**	**<.001**	**0.942**	**<.001**	0.441	.102	**0.870**	**<.001**	**0.871**	**.001**
Item 16	**-1.119**	**<.001**	0.281	.059	**1.552**	**<.001**	-0.123	.506	**0.876**	**<.001**
Item 17	0.154	.380	**0.564**	**<.001**	0.357	.161	**0.277**	**.050**	**0.802**	**.002**
Item 18	**0.669**	**<.001**	-0.312	.083	0.465	.087	0.045	.790	**0.798**	**.003**
Item 19	0.197	.230	**-0.570**	**.003**	**0.543**	**.031**			0.457	.095
Item 20	**-0.831**	**<.001**			**0.536**	**.034**			**1.131**	**<.001**
Item 21	0.183	.202			**1.124**	**<.001**				
Item 22	**0.528**	**<.001**			0.405	.114				
Item 23	**0.884**	**<.001**			0.292	.262				
Item 24	0.073	.641			**2.197**	**<.001**				
Item 25	0.178	.264			0.202	.424				
Item 26	**-0.451**	**.005**			**1.539**	**<.001**				
Item 27	**0.926**	**<.001**			**1.140**	**.021**				
Item 28	0.315	.085			0.338	.262				
Item 29	**0.346**	**.028**								
Item 30	0.237	.065								
Item 31	**1.021**	**<.001**								
Item 32	**0.324**	**.013**								
Item 33	**0.520**	**.007**								

Note. Diff = difference between samples in the level of ability required to obtain half the maximal score on the item, expressed in standard deviations of ability, with a positive sign indicating higher required ability for the 2019 sample than the 1999 reference sample ; *p* = corresponding *p*-value, as obtained with the item parameter replication (IPR) method ; NC = item parameters not computable due to an insufficient number of participants failing the item. Items with significant DIF are in bold.

Table 6

Differential test functioning for the five subtests involving culture-based declarative knowledge

Subtest	Ability required for an intermediate score (2019)	Predicted score for an average ability (2019)	Approximate standard score (2019)	Difference between WAIS-III and WAIS-IV
Vocabulary	+ 0.21	- 2.30	-1	-1.55
Similarities	+ 0.14	- 0.74	-0.5	-1.47
Information	+ 0.65	- 2.61	-1	-1.10
Comprehension	+ 0.18	- 0.48	-0.5	-1.12
Arithmetic	+ 0.55	- 1.87	-1	-0.68

Note. The first column displays the level of ability (standardized) required to obtain half the maximum number of points on the subtest for the 2019 sample, compared to the 1999 sample; the second column displays the predicted raw score obtained with an average level of ability for the 2019 sample, compared to the 1999 sample; the third column displays a coarse estimate of the predicted standard score obtained with an average ability for the 2019 sample; the fourth column displays the difference of average standard scores observed between the WAIS-III and WAIS-IV, combining the Study 1 and Study 2 datasets, for comparison.

3.3 Discussion

The results of Study 2 replicated the conclusions of Study 1: there was a significant difference of FSIQ between WAIS-III and WAIS-IV, but it was exclusively driven by the five subtests depending on *Gc*; no difference appeared for *Gv* or *Gf* subtests. The same was true when combining the two datasets, and Bayesian evidence was clearly in favor of the null for *Gv* and *Gf* subtests, in line with a difference on subtests assessing culture-dependent declarative knowledge but incompatible with the existence of a general decline of intelligence or a negative Flynn effect.

Further analyses investigating DIF indicated significant bias against a recent sample for many items of the five *Gc* subtests of the WAIS-III. DIF was relatively limited for the Similarities and Comprehension subtests, where item parameters could not be estimated for many items which lacked discriminating power in the current sample; but it was clearly apparent for the three other subtests where more items were available. Examining item content suggested a number of things that were comparatively less well known in the recent sample - such as the names of famous people, topics of civic education, and how to perform a mental division. A rough estimation of summed bias at the test level suggested that this DIF was sufficient to account for most or all of the observed difference between WAIS-III and WAIS-IV for the various subtests. In other words, the lower performance on subtests depending on culture-based declarative knowledge was largely attributable, not to a lower ability in the recent sample, but to items being comparatively more difficult for a recent sample than for the normative sample (for an equal level of ability).

One discrepancy with the results of Study 1 was the lack of a difference for Block Design, but since this difference existed only for one of four subgroups in Study 1, the significant difference in the other dataset may have been a spurious effect partly caused by the particular sampling with discrete age groups. Another discrepancy was the significant difference for the Similarities subtest observed in Study 2, which is in line with our hypothesis of a specific difference for *Gc* subtests, but which had not appeared in Study 1. There does not seem to be an obvious explanation for this discrepancy, though we note that Similarities was also the subtest with the least DIF, which is more compatible with the non-significant difference found in Study 1. The third discrepancy was the finding of an *increase* in performance for the Symbol Search subtest, a measure of *Gs*; although processing speed may not be the best index of general intelligence, this is even more incompatible with a general decline of intellectual ability. There were significant procedural differences between the WAIS-III and WAIS-IV versions of Symbol Search, but it is unclear how this would explain the observed difference; this result was also more unstable, with Bayesian analyses failing to find a substantial difference. All three discrepancies between Studies 1 and 2 may represent random variation due to limited sample sizes.

Descriptive statistics indicated that average ability was higher than the WAIS-IV norm in this sample (Table 3; average FSIQ = 111.16). This might be a reflection of a (positive) Flynn effect (with the WAIS-IV normative sample being almost ten years old), but we think it much more likely that this was a sampling bias due to recruitment by word of mouth as performed by people with a university level of education. Consistent with this possibility, this superior performance mostly appeared for the VCI which is the most sensitive to cultural level, whereas averages were close to 100 for the other indices (see Table 3). Importantly, this high average ability did not bias the analyses: there was no ceiling effect, and scores were still in the normal range (the lowest FSIQ was 79), had adequate dispersion and were not skewed (all skewness coefficients were below $|1|$). This point was also accounted for in analyses of DIF, where the normative sample was trimmed to match the composition of the current sample (see section 3.1.3), and where we used item response theory which estimates item parameters using subjects of all available ability levels.

The limitation of this study was its relatively low sample size. For the first analysis testing which subtests demonstrated a difference between WAIS-III and WAIS-IV, this was not a major problem, given that Study 2 corroborated Study 1. Combining the datasets for Study 1 and Study 2 also partly answered this issue, although this was not a perfect strategy either (with the two samples being separated by a little under 10 years, even though their conclusions converged). However, sample size was still low for IRT-based analyses. This issue was attenuated as much as possible by comparing item parameters for DIF analyses to the much larger standardization sample for the WAIS-III retrieved from the publisher (*n* = 767 after matching for ability), and by using scores on the rest of the WAIS-III as an index of subject ability, which made parameter estimation considerably simpler; but estimates presented in Table 5 may be relatively unstable. This is all the more true that the WAIS is not ideal to test for DIF (due in particular to the presence of sequential dependencies between items, with a subtest being interrupted when subjects fail several items in a row; this could especially bias item parameters for the final items of a subtest). Given data collection constraints in France, it would be difficult however to collect a more exhaustive sample.

4. Could the decline be biological ? Revisiting Woodley of Menie & Dunkel (2015)

Woodley of Menie and Dunkel (2015) argued that the results of Dutton and Lynn (2015) indicated a biological origin for the decline of performance between WAIS-III and WAIS-IV in French subjects. This conclusion would be irreconcilable with our finding that the decline is due to differential item functioning related to cultural changes: how do their arguments square with the present results? Their conclusion was based on two results. The first is that the amount of decline on a subtest was correlated with its *g*-loading (a Jensen effect), which they take to indicate biological load; they reported *r* = .83 for this effect. The second is that the amount of decline on a subtest was correlated with its biological load, as estimated by the average of five variables: its *g*-loading, its heritability, its correlation with number of children (an estimate of dysgenics), its correlation with reaction times, and the amount of decline itself; the reported correlation was $r = .72$. Both arguments, however, raise a number of serious statistical and conceptual concerns.

4.1 Statistical and Methodological Issues

First, the correlations of performance decline with both biological load and *g*-loadings are substantial overestimates, and both are actually non-significant at the .05 level. It may seem surprising that the reported correlation of $r = .72$ with biological load is not statistically significant: this is because the authors computed the correlation between performance decline and a composite of five variables including itself. Of course, the correlation of A with a composite (A+B+C+D+E) is artificially inflated. With purely random data, such a composite will tend to share one fifth of its variance ($r = .45$) with each of the five variables from which it is computed, assuming that these variables are uncorrelated. On average, this leads to substantial overestimation of the correlation between the composite and its component variables.

This overestimation requires an adjusted significance test. A Monte-Carlo simulation (replicating the same analysis 100.000 times, using randomly generated data for performance decline; the corresponding code is available as supplemental material) indicated that the *p*-value for the observed $r = .72$ was actually $p = .144$. The critical value for a correlation significantly greater than chance at the .05 level in such an analysis - correlating a variable with a composite of five variables including itself - would have been $r = .81$. Computing the correlation between performance decline, and the same composite including the four indices of biological load but excluding performance decline, gave a fairer $r = .53$, $p = .147$. This correlation was partly driven by *g*-loadings, the only one of the four indices of biological load to correlate significantly with performance decline.

Turning next to the correlation with *g*-loadings (the Jensen effect), the reported $r = .83$ was obtained only after applying an extreme level of correction: to account for sampling error, restriction of range, unreliability and lack of psychometric validity, the raw correlation of $r = .38$, $p = .223$ was multiplied by 2.19. There are three problems with this. The first problem is conceptual: it is well-known that applying this sort of correction can magnify spurious effects in the data (e.g. Winne & Belfry, 1982; to illustrate, a raw correlation of .46 would have yielded a corrected correlation greater than 1). This is especially true when multiple corrections are applied: each correction factor may be imperfectly estimated (and in this case, three out of four correction factors were estimated from different versions of the WAIS or different tests altogether; for another issue related to the independence of scores and correction factors, see Winne & Belfry, 1982). The risk of magnifying spurious effects is especially present when these corrections are large and when they are applied to a small dataset of $N = 12$ (in which graphing the data reveals significant departures from normality). All of this can lead to highly unstable estimates.

The second problem is with the correction for restriction of range. It is not entirely clear whether this correction, which inflated the correlation by about 50%, was justified in this case (this approach makes assumptions about how the data were selected, and what the performance decline would have been with a different version of the WAIS), and besides, it was applied incorrectly. The authors divided the observed correlation by the ratio of the restricted and unrestricted standard deviations, whereas the actual correction is a bit more complex (see e.g. Stauffer & Mendoza, 2001; Wiberg & Sundström, 2009). Using the correct formula gives $r = .74$ instead of $r = .83$.

The third problem is that the authors performed statistical inference on the corrected correlation as if it had been uncorrected; but inflating the correlation with a correction also inflates the risk of error, which also requires an adjusted significance test. Conceptually, this is related to the risk of magnifying spurious effects. A number of authors have attempted to derive appropriate significance tests for a corrected correlation coefficient (e.g. Hakstian et al., 1988; Raju & Brand, 2003), but ultimately the most common solution is to adjust the confidence interval for the corrected correlation by the same correction factor (see Charles, 2005). In the present case, this gives $r = .83$, 95% CI [-0.60, 2.26], or with the appropriate correction for range restriction, $r = .74$, 95% CI [-0.53, 2.01]. It is clear from these results that the reported Jensen effect is both non-significant (the confidence interval includes zero) and highly unstable (with plausible values or *r* ranging between -0.5 and 2). More complex solutions give similar results (e.g. using Raju & Brand, 2003, gives $r = .83$, $Z = 1.22$, $p = .222$).

Lastly, even the non-significant uncorrected correlation of $r = .38$ may be an overestimate. A simple way to demonstrate this is to recalculate the Jensen effect using the *g*-loadings computed, not from the WAIS-III, but from the WAIS-IV (there is no reason to prefer one or the other in the context of a comparison between WAIS-III and WAIS-IV): in this case, the raw Jensen effect was only $r = .29$, $p = .361$ (the correlation is identical whether using the Study 1 dataset like the original authors, or the combined datasets of Study 1 and 2). In short, if a Jensen effect exists here, its magnitude is probably limited.

4.2 Conceptual Issues: Biological Load and Cultural Load

The second and bigger issue is conceptual: most of the indices of biological load used by the authors are questionable. For example, reaction times are not just "biological", given that they are influenced by a host of non-biological variables such as familiarity with test materials, motivation, attention, response strategies, understanding instructions, etc. (see Detterman, 1987; Nettelbeck, 1998; Flynn, 2009a), all of which depend on culture and environment. The interpretation of heritability estimates poses well-known issues and has been criticized elsewhere (e.g. Dickens & Flynn, 2001; Kan et al., 2013). In the case of the

Jensen effect, there is no indication that it is biological: if something correlates with *g*-loadings, it does not mean that this something is biological. At its core, the *g* loading of a subtest can be viewed as an index of its cognitive complexity (Flynn, 2013; see also Gottfredson, 2016): a subtest with a higher *g*-loading is more complex, which means it may be affected to a greater extent not only by biological factors, but also by cultural factors (see especially Dickens & Flynn, 2001; Kan et al., 2013). This is also true for heritability coefficients. In other words, it is actually impossible to directly disentangle biological and environmental causes based on the measures used by the authors. This is summarized in the findings of Kan and colleagues (2013), who found that the *g*-loadings and heritability coefficients of a test were substantially correlated with its cultural load ($r = .83$ and $r = .40$ respectively).

This logically invites a new analysis: Woodley of Menie and Dunkel (2015) attempted to find a correlation between performance decline and *biological* load, but they failed to search for a correlation between performance decline and *cultural* load. We performed this analysis, using the estimate of performance decline computed on the combined Study 1 and Study 2 datasets (see partial eta squared effect sizes in Table 4; the results were very similar when considering only the Study 1 dataset). We used three different measures of cultural load: the subjective influence of culture on a subtest as assessed by expert consensus (retrieved from Kan et al., 2013); the average proportion of items that needed to be changed when adapting the WAIS-III to a different country (retrieved from Georgas et al., 2003; see also Kan et al., 2013)[3]; and the effect size of socio-economic level on performance (expressed

[3] Subjective cultural load and the average proportion of items that needed to be changed were not available for three subtests: Matrix Reasoning, Letter-Number Sequencing, and Symbol Search. Their subjective loads were fixed to zero based on similar tests (Kan et al., 2013); and their proportion of items changed was also fixed to zero based on the fact that no items were changed in the French adaptation of the WAIS. Excluding these three subtests from the analysis did not change the results displayed in Table 7 (it decreased the correlation between performance decline and composite cultural load to $r = .94$).

as eta squared, which we computed directly from the French WAIS-III normative sample). These three measures were also standardized and averaged to yield a composite cultural load index.

The correlations between performance decline and cultural load are displayed in Table 7. Performance decline on a subtest correlated near unity with its cultural load: the correlations were between .85 and .93 for the three measures of cultural load considered separately, and reached .95 for the composite cultural load. All correlations were significant at the .001 level. Note that these are the raw bivariate correlations, without correction for attenuation or multi-vector analysis (correcting subtest-level estimates for their reliability, as in the second analysis of Woodley of Menie & Dunkel, 2015, changed the correlation estimates by less than .02). Thus, the cultural load of a subtest accounted for virtually all of its decline.

In our opinion, the fact that performance decline on a subtest correlated .95 with its cultural load, and that heritability and dysgenics both had non-significant correlations with performance decline, clearly supports the idea that the source of the decline is cultural rather than biological. This converges with the results of Studies 1 and 2, which showed that there was no decline for the subtests that did not load on *Gc*, and that subjects in the 2019 sample performed lower on these subtests even when they had equal scores on the rest of the WAIS.

Note that these results are not sufficient to claim that the decline is exclusively cultural: as noted above, all coefficients - whether of cultural load or biological load - reflect a mix of cultural and biological influences. For example, it could still be the case that a biological factor influenced culturally-loaded subtests due to a genetically-driven cultural decrease (Dutton et al., 2017). Although this is a possibility, such a hypothesis would be almost impossible to falsify (if cultural changes exclusively reflect biological changes, there is no way to show an influence of something other than biology), and the question would remain of what biological change could have driven the decline. As noted by Woodley of Menie and Dunkel (2015), substantial genetic changes are unlikely over such a short timeframe, and their proposed explanation of immigration would be very difficult to defend in this particular case: there did not seem to be large differences in racial composition of the two WAIS samples, and the share of immigrants in the French population has increased by about 2 percentage points between 1999 and 2019, which seems much too small to create the large differences reported here (e.g. Figure 2b).

Table 7

Performance decline and cultural load on a subtest, and their correlation

Subtest	Performance decline	Cultural load (subjective)	Cultural load (changed items)	Cultural load (socio-economic)	Composite cultural load
Vocabulary	0,32	1	0,351	0,299	1,607
Similarities	0,22	1	0,091	0,291	0,757
Information	0,20	1	0,221	0,337	1,403
Comprehension	0,19	1	0,151	0,304	1,010
Picture completion	0,00	0	0,031	0,162	-0,789
Block design	0,01	0	0,011	0,183	-0,742
Matrix reasoning	0,00	0	0	0,228	-0,538
Arithmetic	0,08	1	0,081	0,256	0,540
Digit span	0,00	0	0,001	0,167	-0,860
Letter-number sequencing	0,00	0	0	0,210	-0,633
Digit symbol coding	0,01	0	0,002	0,137	-1,016
Symbol search	-0,03	0	0	0,190	-0,739
	Correlation with performance decline	.89 ***	.93 ***	.85 ***	.95 ***

Note. *** $p < .001$. Performance declines computed from the combined Study 1 and Study 2 datasets; subjective cultural loads are from Kan et al. (2013); proportions of items changed in cross-cultural adaptations are from Georgas et al. (2003) and Kan et al. (2013); the effect of socio-economic level was computed from the WAIS-III normative dataset; composite cultural load is the average of the three cultural load estimates after standardization.

5. General Discussion

The results of both Study 1 and Study 2 unambiguously indicated that there was no negative Flynn effect in France, in the sense of a general decrease of intelligence or a decrease in the ability to perform logical reasoning: there were no reliable differences between WAIS-III and WAIS-IV for any of the subtests reflecting visuo-spatial reasoning (*Gf* and *Gv*), or working memory and processing speed (*Gsm* and *Gs*), and which were based on abstract materials. We did find lower total performance on the WAIS-III for a recent sample, but contrary to the classic Flynn effect, this difference between cohorts was exclusively driven by the five subtests involving *Gc* - acquired declarative knowledge tied to a specific cultural setting.

When considered under the angle of item content, it appeared that this decrease on subtests involving declarative knowledge largely reflected, not an actual decrease of ability, but measurement bias due to differences of item difficulty for samples collected at different dates. All in all, in the five subtests demonstrating a decline, about one fourth of items were comparatively more difficult for the 2019 sample than for the 1999 sample for an equal level of ability. These differences could be traced down to a few specific skills. All but one of the Information items that were biased against a recent sample related to the names of famous people, and biased Comprehension items were all related to civic education; interestingly, the test publisher decided to practically eliminate both topics from the WAIS-IV. All but one of the biased Arithmetic items required computing mental division or proportions. For Vocabulary, the negative net effect of bias was partly compensated by the fact that some words were easier in the recent sample, more consistent with a change in language frequency patterns than with an absolute decrease in vocabulary skills. In all cases, these increases in item difficulty for a recent sample could be attributed to environmental changes in school programs, topics covered by the media, and other societal evolutions.

The fact that the performance decrease on a subtest correlated at .95 with its cultural load confirms this conclusion and runs counter to the interpretation that the observed decline is caused by biological factors (Woodley of Menie & Dunkel, 2015). This does not completely rule out biological factors, as cultural loads are not pure indicators of cultural influences: a possible alternative interpretation, as suggested by Edward Dutton and Woodley of Menie, is that a genetic decrease in fluid reasoning could negatively affect the culture of a country, in turn reverberating on *Gc* subtests (see Dutton et al., 2017; this is a variant of investment theory and of explanations assuming genotype-environment covariance; e.g. Kan et al., 2013). However, this idea would be almost impossible to falsify, and it would be difficult to reconcile with the facts that the correlation with heritability was non-significant and that there was no decline at all for the *Gf* and *Gv* subtests, which tend to have high heritability (e.g. Kan et al., 2013; Rijsdijk et al., 2002; van Leuuwen et al., 2008), and which would be expected to decrease before effects on *Gc* could be observable. There is also a lack of plausible biological mechanisms that could create such a large decline in the dataset in such a short timeframe. All this converges to suggest a role of cultural changes as the most parsimonious interpretation of the data.

In short, the conclusion that can be drawn from a comparison of WAIS-III and WAIS-IV is that over the last two decades, there has been no decline of reasoning abilities in the French population, but there has been an average decrease in a limited range of cultural knowledge (essentially related to using infrequent vocabulary words, knowing the names of famous people, discussing civic education and performing mental division), which biases performance on older items. In other words, the data do indicate a lower average *performance* on the WAIS-III in the more recent sample, in line with Dutton and Lynn (2015)'s results, but a more fine-grained analysis contradicts their interpretation of a general decrease of *intelligence* in France. In the terms of a hierarchical model of intelligence (Wicherts, 2007), there appears to be no decrease in latent ability at the first level of *g*; there is a decrease at the second level of broad abilities, but only for *Gc*; and this decrease seems essentially due to cultural changes creating measurement bias at the fourth level composed of performance for specific items.

This pattern is entirely distinct from the Flynn effect, which represents an increase in general intelligence, and especially in *Gf* performance, accompanied by much smaller changes on *Gc* (Pietschnig & Voracek, 2015). Hence it is our conviction that this pattern reflects substantially different mechanisms and cannot reasonably be labeled a "negative Flynn effect", without extending the definition of the Flynn effect to the point where any difference between cohorts could be called a "Flynn effect" and where it would no longer be useful as a heuristic concept. This point is compounded by the fact that the difference reflected item-related measurement bias, rather than an actual change of ability. To quote Flynn (2009a): "*Are IQ gains 'cultural bias' ? We must distinguish between cultural trends that render neutral content more familiar and cultural trends that really raise the level of cognitive skills. If the spread of the scientific ethos has made people capable of using logic to attack a wider range of problems, that is a real gain in cognitive skills. If no one has taken the trouble to update the words on a vocabulary test to eliminate those that have gone out of everyday usage, then an apparent score loss is ersatz.*" The current pattern is clearly ersatz: "ersatz effect" may be a better name than "negative Flynn effect".

There are two possible interpretations to the ersatz difference observed here. On one hand, this decline could be restricted to areas covered by the WAIS-III, and could be compensated by increases in other areas: in other words, the 2019 sample may possess *different knowledge*, but not *less knowledge* than the 1999 sample. On the other hand, this might represent a real decline and a cause for concern: results of the large-scale PISA surveys (performed on about 7.000 pupils) routinely point to significant inequalities in the academic skills of French pupils, and their average level of mathematics performance has declined since the early 2000s (e.g. OECD, 2019). It is impossible to adjudicate between these two possibilities (which would require having the 1999 sample perform the WAIS-IV), but even if there were an actual decrease in average knowledge, this conclusion would be significantly less bleak than the picture of a biologically-driven intelligence decrease painted by Dutton and Lynn (2015), and would highlight possible shortfalls of the French educational system (see also Blair et al., 2005) rather than the downward trajectory of a population becoming less and less intelligent.

This conclusion is in line with a tradition of studies attributing fluctuations of intelligence scores to methodological biases, especially as they relate to [cultural] item content (e.g. Beaujean & Osterlind, 2008; Beaujean & Sheng, 2010; Kaufman, 2010; Nugent, 2006; Pietschnig et al., 2013; Rodgers, 1998; Weiss et al., 2015). As an example, Flieller (1988) reached the same conclusion in a French dataset over three decades ago; Brand and colleagues (1989) also found a similar result of decreasing scores due to changes of items difficulty, which they illustrated with an understandable decline of the proportion of correct answers for the item "What is a belfry?" between 1961 and 1984. This conclusion is also in line with studies arguing for the role of cultural environment and culture-based knowledge in Flynn-like fluctuations of intelligence over time (e.g. Bratsberg & Rogeberg, 2018). Note that drifts of item difficulty are only one aspect of such cultural changes; changes of test-taking pattern behavior, such as increased guessing, are another example (e.g. Must & Must, 2013; Pietschnig & Voracek, 2013).

Beyond the specific case of average intelligence in France, the current results constitute a reminder that intelligence scores are not pure reflections of intelligence and have multiple determinants, some of which can be affected by cultural factors that do not reflect intelligence itself. Put otherwise, this is an illustration of the principle that performance can differ between groups of subjects without representing a true difference of ability (Beaujean & Osterlind, 2008; Beaujean & Sheng, 2010). This is a well-known bias of cross-country comparisons, where test performance can be markedly lower in a culture for which the test was not designed (e.g. Cockcroft et al., 2015; Greenfield, 1997; Van de Vijver, 2016). In other words, this principle generalizes to all comparisons between samples, not just intelligence fluctuations over time: investigators should be skeptical of the origin of between-group differences whenever cultural content is involved. This also applies to clinical psychologists using intelligence tests to compare patients from specific cultural groups to a (culturally different) normative sample.

Seven major recommendations for cross-sample comparisons can be derived from the current results:

1) comparisons based on validity samples collected by the publishers of Wechsler scales have to be avoided due to uncertainties about sample composition (as already stressed by Zhu and Tulsky, 1999; the distribution of ages in Study 1 as represented in Figure 1 constitutes a stark reminder of this fact);
2) comparisons involving multiple subtests should carefully consider which subtests exactly demonstrate differences, and especially which dimension of intelligence they measure (*Gf* or *Gc*?);
3) comparisons between different samples should never be performed using different tests with substantial differences of item content, if there is a possibility that the items will be differentially affected by cultural variables extraneous to ability itself (Kaufman, 2010; Weiss et al., 2015);

4) even when the same version of a test involving cultural content is used, differences between samples collected at different dates in the same country should be treated as if the past sample were from a different country, due to the possibility of differential item functioning emerging over time;

5) as a consequence, comparisons between samples should primarily rely on tests that involve as little contribution of culture-based declarative knowledge as possible, such as Raven's matrices (e.g. Flynn, 2009b);

6) when only tests requiring culture-based declarative knowledge are available, differences should necessarily be interpreted taking into account possible measurement bias. The issue of measurement bias can be considered under the prism of IRT as a way to separate item parameters from ability estimates and test for DIF, and/or using multigroup confirmatory factor analyses as a way to more accurately specify at which level of a hierarchical model of intelligence samples actually differ (Wicherts et al., 2004);

7) lastly, and as exemplified by the pattern of correlations between performance decline, heritability and *g*-loadings, and cultural load, no conclusions about the biological origin of between-group differences in test scores can be drawn without also testing the role of cultural factors.

Acknowledgements

The authors thank Pearson and the ECPA (*les Editions du Centre de Psychologie Appliquée*) for authorizing access to the WAIS-III and WAIS-IV normative data.

References

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement*, *29*(1), 67–91. doi:10.1111/j.1745-3984.1992.tb00368.x

Ban, J. C., Hanson, B. A., Wang, T., Yi, Q., & Harris, D., J. (2001) A comparative study of on-line pretest item calibration/scaling methods in computerized adaptive testing. *Journal of Educational Measurement*, *38*(3), 191-212. doi:10.1111/j.1745-3984.2001.tb01123.x

Baxendale, S. (2011). IQ and ability across the adult life span. *Applied Neuropsychology*, *18*(3), 164–167. doi:10.1080/09084282.2011.595442

Beaujean, A. A., & Osterlind, S. J. (2008). Using item response theory to assess the Flynn effect in the National Longitudinal Study of Youth 79 Children and Young Adults data. *Intelligence*, *36*(5), 455–463. doi:10.1016/j.intell.2007.10.004

Beaujean, A., & Sheng, Y. (2010). Examining the Flynn effect in the general social survey vocabulary test using item response theory. *Personality and Individual Differences*, *48*(3), 294–298. doi:10.1016/j.paid.2009.10.019

Beaujean, A., & Sheng, Y. (2014). Assessing the Flynn Effect in the Wechsler scales. *Journal of Individual Differences*, *35*(2), 63–78. doi:10.1027/1614-0001/a000128

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289-300. doi:j.2517-6161.1995.tb02031.x

Benson, N., Hulac, D. M., & Kranzler, J. H. (2010). Independent examination of the Wechsler Adult Intelligence Scale—Fourth Edition (WAIS-IV): What does the WAIS-IV measure? *Psychological Assessment*, *22*(1), 121–130. doi:10.1037/a0017767

Birnbaum, A. (1968). Some Latent Trait Models and Their Use in Inferring an Examinee's Ability. In F. M. Lord and M. R. Novick (Eds.), *Statistical Theories of Mental Test Scores* (pp. 397-479). Addison-Wesley.

Blair, C., Gamson, D., Thorne, S., & Baker, D. (2005). Rising mean IQ: Cognitive demand of mathematics education for young children, population exposure to formal schooling, and the neurobiology of the prefrontal cortex. *Intelligence*, *33*(1), 93–106. doi:10.1016/j.intell.2004.07.008

Brand, C. R., Freshwater, S., & Dockrell, W. B. (1989). Has there been a "massive" rise in IQ levels in the West? Evidence from Scottish children. *The Irish Journal of Psychology*, *10*(3), 388–393. doi:10.1080/03033910.1989.10557756

Bratsberg, B., & Rogeberg, O. (2018). Flynn effect and its reversal are both environmentally caused. *PNAS Proceedings of the National Academy of Sciences of the United States of America*, *115*(26), 6674–6678. doi:10.1073/pnas.1718793115

Canivez, G. L., & Watkins, M. W. (2010). Investigation of the factor structure of the Wechsler Adult Intelligence Scale—Fourth Edition (WAIS–IV): Exploratory and higher order factor analyses. *Psychological Assessment*, *22*(4), 827–836. doi:10.1037/a0020429

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review*, *97*, 404–431. http://dx.doi.org/10.1037/0033-295X.97.3.404

Cervantes, V. H. (2012). On using the item parameter replication (IPR) approach for power calculation of the noncompensatory DIF (NCDIF) index. In C. Arce and G. Seoane (Eds.), *5th European Congress of Methodology – Book of Abstracts* (pp. 206–207). Universidade de Santiago de Compostela.

Cervantes, V. H. (2017a). DFIT: An R package for Raju's differential functioning of items and tests framework. *Journal of Statistical Software*, *76*(5), 1-24. doi:10.18637/jss.v076.i05

Cervantes, V. H. (2017b). *DFIT: An R package for the differential functioning of items and tests framework*. Instituto Colombiano para la Evaluación de la Educación [ICFES], Bogotá, Colombia. R package version 1.0-3. https://CRAN.R-project.org/package=DFIT.

Charles, E. P. (2005). The Correction for Attenuation Due to Measurement Error: Clarifying Concepts and Creating Confidence Sets. *Psychological Methods*, *10*(2), 206–226. doi:10.1037/1082-989X.10.2.206

Choi, S. W. (2016). lordif: Logistic Ordinal Regression Differential Item Functioning using IRT. R package version 0.3-3. https://CRAN.R-project.org/package=lordif

Clark, P. C., & LaHuis, D. M. (2012). An examination of power and Type I errors for two differential item functioning indices using the graded response model. *Organizational Research Methods*, *15*(2), 229–246. doi:10.1177/1094428111403815.

Cockcroft, K., Alloway, T., Copello, E., & Milligan, R. (2015). A cross-cultural comparison between South African and British students on the Wechsler Adult Intelligence Scales Third Edition (WAIS-III). *Frontiers in Psychology*, *6*(297). doi:10.3389/fpsyg.2015.00297

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Routledge.

Detterman, D. K. (1987). What does reaction time tell us about intelligence? In P. A. Vernon (Ed.), *Speed of information-processing and intelligence* (pp. 177–200). Ablex Publishing.

Dickens, W. T., & Flynn, J. R. (2001). Heritability estimates versus large environmental effects: The IQ paradox resolved. *Psychological Review*, *108*(2), 346–369. doi:10.1037/0033-295X.108.2.346

Dori, G. A., & Chelune, G. J. (2004). Education-Stratified Base-Rate Information on Discrepancy Scores Within and Between the Wechsler Adult Intelligence Scale-Third

Edition and the Wechsler Memory Scale-Third Edition. *Psychological Assessment*, *16*(2), 146–154. doi:10.1037/1040-3590.16.2.146

Dutton, E., Bakhiet, S. F., Essa, Y. A. S., Blahmar, T. A., & Hakami, S. M. A. (2017). A Negative Flynn Effect in Kuwait: The same effect as in Europe but with seemingly different causes. *Personality and Individual Differences*, *114*, 69–72. doi:10.1016/j.paid.2017.03.060

Dutton, E., & Lynn, R. (2013). A negative Flynn effect in Finland, 1997–2009. *Intelligence*, *41*(6), 817–820. doi:10.1016/j.intell.2013.05.008

Dutton, E., & Lynn, R. (2015). A negative Flynn Effect in France, 1999 to 2008–9. *Intelligence*, *51*, 67–70. doi:10.1016/j.intell.2015.05.005

Dutton, E., van der Linden, D., & Lynn, R. (2016). The negative Flynn Effect: A systematic literature review. *Intelligence*, *59*, 163–169. doi:10.1016/j.intell.2016.10.002

Flanagan, D. P. (2000). Wechsler-based CHC cross-battery assessment and reading achievement: Strengthening the validity of interpretations drawn from Wechsler test scores. *School Psychology Quarterly*, *15*(3), 295–329. doi:10.1037/h0088789

Flanagan, D. P., Alfonso, V. C., & Reynolds, M. R. (2013). Broad and narrow CHC abilities measured and not measured by the Wechsler Scales: Moving beyond within-battery factor analysis. *Journal of Psychoeducational Assessment*, *31*(2), 202–223. doi:10.1177/0734282913478047

Flieller, A. (1988). Application du modèle de Rasch à un problème de comparaison de générations [Applications of the Rasch model to a problem of intergenerational comparison]. *Bulletin de Psychologie*, *42*(388), 86-91.

Flynn, J. R. (1984). The mean IQ of Americans: Massive gains 1932 to 1978. *Psychological Bulletin*, 95(1), 29–51. doi:10.1037/0033-2909.95.1.29

Flynn, J. R. (1998a). IQ gains over time: Toward finding the causes. In U. Neisser (Ed.), The rising curve: Long-term gains in IQ and related measures (pp. 25–66). American Psychological Association. doi:10.1037/10270-001

Flynn, J. R. (1998b). WAIS-III and WISC-III gains in the United States from 1972 to 1995: How to compensate for obsolete norms. *Perceptual and Motor Skills*, *86*(3, Pt 2), 1231–1239. doi:10.2466/pms.1998.86.3c.1231

Flynn, J. R. (2009a). *What is intelligence?* Cambridge University Press.

Flynn, J. R. (2009b). Requiem for nutrition as the cause of IQ gains: Raven’s gains in Britain 1938–2008. *Economics and Human Biology*, *7*, 18–27. doi:10.1016/j.ehb.2009.01.009

Flynn, J. R. (2013). The “Flynn Effect” and Flynn’s paradox. *Intelligence*, *41*(6), 851–857. doi:10.1016/j.intell.2013.06.014

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., & Hothorn, T. (2019). *mvtnorm: Multivariate normal and t distributions*. R package version 1.0-11. http://CRAN.R-project.org/package=mvtnorm

Georgas, J., van de Vijver, F. J. R., Weiss, L. G., & Saklofske, D. H. (2003). A cross-cultural analysis of the WISC-III. In J. Georgas, L. G. Weiss, F. J. R. van de Vijver, & D. H. Saklofske (Eds.), *Culture and children's intelligence: Cross-cultural analysis of the WISC-III* (pp. 277–313). Academic Press. doi:10.1016/B978-012280055-9/50021-7

Golay, P., & Lecerf, T. (2011). Orthogonal higher order structure and confirmatory factor analysis of the French Wechsler Adult Intelligence Scale (WAIS-III). *Psychological Assessment*, *23*(1), 143–152. doi:10.1037/a0021230

Gottfredson, L. S. (2016). A g theorist on why Kovacs and Conway's process overlap theory amplifies, not opposes, g theory. *Psychological Inquiry*, *27*(3), 210–217. doi:10.1080/1047840X.2016.1203232

Greenfield, P. (1997). You can't take it with you: Why ability assessments don't cross cultures. *American Psychologist*, *52*(10), 1115-1124. doi:10.1037/0003-066X.52.10.1115

Grégoire, J. (1993). Intelligence et vieillissement au WAIS-R Une analyse transversale de l'échantillon d'étalonnage français avec contrôle du niveau scolaire [Intelligence and aging: A cross-sectional analysis of the French standardization sample of the WAIS-R with educational level controlled]. *L'Année Psychologique*, *93*(3), 379–400. doi:10.3406/psy.1993.28701

Grégoire, J., Daniel, M., Llorente, A. M., & Weiss, L. C. (2016). The Flynn effect and its clinical implications. In L. G. Weiss, D. H. Saklofske, J. A. Holdnack, & A. Prifitera (Eds.), WISC-*V assessment and interpretation: Scientist-practitioner perspectives* (pp. 187–212). Elsevier Academic Press. doi:10.1016/B978-0-12-404697-9.00006-6

Hakstian, A. R., Schroeder, M. L., & Rogers, W. T. (1988). Inferential procedures for correlation coefficients corrected for attenuation. *Psychometrika*, *53*(1), 27–43. doi:10.1007/BF02294192

Heaton, R. K., Taylor, M. J., & Manly, J. (2003). Demographic effects and use of demographically corrected norms with the WAIS-III and WMS-III. In D. S. Tulsky, D. H. Saklofske, G. J. Chelune, R. K. Heaton, R. J. Ivnik, R. Bornstein, A. Prifitera, & M. F. Ledbetter (Eds.), *Clinical interpretation of the WAIS-III and WMS-III* (pp. 181–210). Academic Press. doi:10.1016/B978-012703570-3/50010-9

Holdnack, J. A., Drozdick, L. W., Weiss, L. G., & Iverson, G. L. (2013). *WAIS-IV, WMS-IV, and ACS: Advanced clinical interpretation*. Elsevier Academic Press.

Jensen, A. R. (1994). Phlogiston, animal magnetism, and intelligence. In D. K. Detterman (Ed.), *Current topics in human intelligence, Vol. 4: Theories of intelligence* (pp. 257-284). Ablex.

Kan, K.-J., Wicherts, J. M., Dolan, C. V., & van der Maas, H. L. J. (2013). On the nature and nurture of intelligence and specific cognitive abilities: The more heritable, the more culture dependent. *Psychological Science*, *24*(12), 2420–2428. doi:10.1177/0956797613493292

Kaufman, A. S., Reynolds, C. R., & McLean, J. E. (1989). Age and WAIS—R intelligence in a national sample of adults in the 20- to 74-year age range: A cross-sectional analysis with educational level controlled. *Intelligence*, *13*(3), 235–253. doi:10.1016/0160-2896(89)90020-2

Kaufman, A. S. (2010). "In what way are apples and oranges alike?" A critique of Flynn's interpretation of the Flynn effect. *Journal of Psychoeducational Assessment*, *28*(5), 382–398. doi:10.1177/0734282910373346

Keith, T. Z., Fine, J. G., Taub, G. E., Reynolds, M. R., & Kranzler, J. H. (2006). Higher order, multisample, confirmatory factor analysis of the Wechsler Intelligence Scale for Children--Fourth Edition: What does it measure? *School Psychology Review*, *35*(1), 108–127.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*. doi:10.3389/fpsyg.2013.00863

Lecerf, T., Golay, P., & Reverte, I. (2012). Scores composites CHC pour la WAIS-IV: Normes francophones [CHC composite scores for the WAIS-IV: French Norms]. *Pratiques Psychologiques*, *18*(4), 401–412. doi:10.1016/j.prps.2012.03.001

Lim, H. (2020). *irtplay: Online item calibration, scoring, and evaluation of model-data fit in Item Response Theory*. R package version 1.4.1. https://CRAN.R-project.org/package=irtplay

Marshalek, B., Lohman, D. F., & Snow, R. E. (1983). The complexity continuum in the radex and hierarchical models of intelligence. *Intelligence*, *7*(2), 107–127. doi:10.1016/0160-2896(83)90023-5

Martinková, P., Drabinová, A., Liaw, Y.-L., Sanders, E. A., McFarland, J. L., & Price, R. M. (2017). Checking equity: Why DIF analysis should be a routine part of developing conceptual assessments. *CBE - Life Sciences Education*, *16*(2), 1-13. doi:10.1187/cbe.16-10-0307

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174. doi:10.1007/BF02296272

McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence*, *37*(1), 1–10. doi:10.1016/j.intell.2008.08.004

Must, O., & Must, A. (2013). Changes in test-taking patterns over time. *Intelligence*, *41*, 791–801. doi:10.1016/j.intell.2013.04.005

Nettelbeck, T. (1998). Jensen's chronometric research: Neither simple nor sufficient but a good place to start. *Intelligence*, *26*(3), 233–241. doi:10.1016/S0160-2896(99)80006-3

Nugent, W. R. (2006). The Comparability of the Standardized Mean Difference Effect Size Across Different Measures of the Same Construct: Measurement Considerations.

Educational and Psychological Measurement, *66*(4), 612–623. doi:10.1177/0013164405284032

OECD (2019), *PISA 2018 results (Volume I): What students know and can do*. PISA, OECD Publishing. doi:10.1787/5f07c754-en

Oshima, T. C., Raju, N. S., & Nanda, A. O. (2006). A new method for assessing the statistical significance in the differential functioning of items and tests (DFIT) framework. *Journal of Educational Measurement*, *43*(1), 1–17. doi:10.1111/j.1745-3984.2006.00001.x.

Pietschnig, J., Tran, U. S., & Voracek, M. (2013). Item-response theory modeling of IQ gains (the Flynn effect) on crystallized intelligence: Rodgers' hypothesis yes, Brand's hypothesis perhaps. *Intelligence*, 41, 791–801. doi:10.1016/j.intell.2013.06.005

Pietschnig, J., & Voracek, M. (2015). One century of global IQ gains: A formal meta-analysis of the Flynn effect (1909–2013). *Perspectives on Psychological Science*, *10*(3), 282–306. doi:10.1177/1745691615577701

R Core Team (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, *53*(4), 495–502. doi:10.1007/bf02294403.

Raju, N. S., & Brand, P. A. (2003). Determining the significance of correlations corrected for unreliability and range restriction. *Applied Psychological Measurement*, *27*(1), 52–71. doi:10.1177/0146621602239476

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Danish Institute for Educational Research.

Rijsdijk, F. V., Vernon, P. A., & Boomsma, D. I. (2002). Application of hierarchical genetic models to Raven and WAIS subtests: A Dutch twin study. *Behavior Genetics*, *32*(3), 199–210. doi:10.1023/a:1016021128949

Rodgers, J. L. (1998). A critique of the Flynn Effect: Massive IQ gains, methodological artifacts, or both? *Intelligence*, *26*(4), 337–356. doi:10.1016/S0160-2896(99)00004-5

Ryan, J. J., Sattler, J. M., & Lopez, S. J. (2000). Age effects on Wechsler Adult Intelligence Scale-III subtests. *Archives of Clinical Neuropsychology*, *15*(4), 311–317. doi:10.1016/S0887-6177(99)00019-0

Rundquist, E. A. (1936). Intelligence test scores and school marks of high school seniors in 1929 and 1933. *School & Society*, *43*, 301–304.

Shayer, M., Ginsburg, D., & Coe, R. (2007). Thirty years on - a large anti-Flynn effect? The Piagetian test Volume & Heaviness norms 1975-2003. *British Journal of Educational Psychology*, *77*(1), 25–41. doi:10.1348/000709906X96987

Stauffer, J. M., & Mendoza, J. L. (2001). The proper sequence for correcting correlation coefficients for range restriction and unreliability. *Psychometrika*, *66*(1), 63–68. doi:10.1007/BF02295732

Stocking, M. L. (1988). *Scale drift in on-line calibration* (Research Rep. 88-28). Educational Testing Service.

Sundet, J. M., Barlaug, D. G., & Torjussen, T. M. (2004). The end of the Flynn effect? A study of secular trends in mean intelligence test scores of Norwegian conscripts during half a century. *Intelligence*, *32*(4), 349–362. doi:10.1016/j.intell.2004.06.004

Teasdale, T. W., & Owen, D. R. (2008). Secular declines in cognitive test scores: A reversal of the Flynn Effect. *Intelligence*, *36*(2), 121–126. doi:10.1016/j.intell.2007.01.007

Trahan, L. H., Stuebing, K. K., Fletcher, J. M., & Hiscock, M. (2014). The Flynn effect: A meta-analysis. *Psychological Bulletin*, *140*(5), 1332–1360. doi:10.1037/a0037173

Van de Vijver, F. J. R. (2016). Assessment in education in multicultural populations. In G. T. L. Brown & L. R. Harris (Eds.), *Handbook of human and social conditions in assessment* (pp. 436-453). Routledge.

van Leeuwen, M., van den Berg, S. M., & Boomsma, D. I. (2008). A twin-family study of general IQ. *Learning and Individual Differences*, *18*(1), 76–88. doi:10.1016/j.lindif.2007.04.006

Ward, C. L., Bergman, M. A., & Hebert, K. R. (2012). WAIS-IV subtest covariance structure: Conceptual and statistical considerations. *Psychological Assessment*, *24*(2), 328-340. doi:10.1037/a0025614

Wechsler, D. (2000). *Manuel de l'Echelle d'Intelligence de Wechsler Pour Adultes - 3ème édition* [Manual for the Wechsler Adult Intelligence Scale - Third Edition]. ECPA.

Wechsler, D. (2011). *Manuel de l'Echelle d'Intelligence de Wechsler Pour Adultes - 4ème édition* [Manual for the Wechsler Adult Intelligence Scale - Fourth Edition]. ECPA par Pearson.

Weiss, L. G., Gregoire, J., & Zhu, J. (2016). Flaws in Flynn effect research with the Wechsler scales. *Journal of Psychoeducational Assessment*, *34*(5), 411–420. doi:10.1177/0734282915621222

Wiberg, M., & Sundström, A. (2009) A comparison of two approaches to correction of restriction of range in correlation analysis. *Practical Assessment, Research, and Evaluation*, *14*(5). doi:10.7275/as0k-tm88

Wicherts, J. M. (2007). *Group differences in intelligence test performance* [Unpublished dissertation]. University of Amsterdam. Retrieved from: https://pure.uva.nl/ws/files/4175964/46967_Wicherts.pdf

Wicherts, J. M., Dolan, C. V., Hessen, D. J., Oosterveld, P., van Baal, G. C. M., Boomsma, D. I., & Span, M. M. (2004). Are intelligence tests measurement invariant over time?

Investigating the nature of the Flynn effect. *Intelligence*, *32*(5), 509–537. doi:10.1016/j.intell.2004.07.002

Winne, P. H., & Belfry, M. J. (1982). Interpretive problems when correcting for attenuation. *Journal of Educational Measurement*, *19*(2), 125–134. doi:10.1111/j.1745-3984.1982.tb00121.x

Woodley of Menie, M. A., & Dunkel, C. S. (2015). In France, are secular IQ losses biologically caused? A comment on Dutton and Lynn (2015). *Intelligence*, *53*, 81–85. doi:10.1016/j.intell.2015.08.009

Zhu, J., & Tulsky, D. S. (1999). Can IQ gain be accurately quantified by a simple difference formula? *Perceptual and Motor Skills*, *88*(3, Pt 2), 1255–1260. doi:10.2466/PMS.88.3.1255-1260

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, *4*(2), 223–233. doi:10.1080/15434300701375832