Flynn effects are biased by differential item functioning over time: A test using overlapping items in Wechsler scales

Corentin Gonthier, Jacques Grégoire

Flynn Effects are Biased by Differential Item Functioning over Time:

A Test using Overlapping Items in Wechsler scales

Corentin Gonthier [1,2] & Jacques Grégoire [3]

[1] Nantes Université, LPPL UR 4638, Nantes, France

[2] Institut Universitaire de France

[3] Université Catholique de Louvain, 1348 Louvain-la-Neuve, Belgique

Correspondence concerning this article should be addressed to Corentin Gonthier, Laboratoire LPPL, Chemin de la Censive du Tertre, BP 81227, 44312 Nantes Cedex 3, France

E-mail: corentin.gonthier@univ-nantes.fr

Abstract

The items of intelligence tests can demonstrate differential item functioning across different groups: cross-sample differences in item difficulty or discrimination, independently of any difference of ability. This is also true of comparisons over time: as the cultural context changes, items may increase or decrease in difficulty. This phenomenon is well-known, but its impact on estimates of the Flynn effect has not been systematically investigated. In the current study, we tested differential item functioning in a subset of 111 items common to consecutive versions of the French WAIS-R (1989), WAIS-III (1999) and/or WAIS-IV (2009), using the three normative samples (total $N = 2979$). Over half the items had significant differential functioning over time, generally becoming more difficult from one version to the next for the same level of ability. The magnitude of differential item functioning tended to be small for each item separately, but the cumulative effect over all items led to underestimating the Flynn effect by about 3 IQ points per decade, a bias close to the expected size of the effect itself. In this case, this bias substantially affected the conclusions, even creating an ersatz negative Flynn effect for the 1999-2009 period, when in fact ability increased (1989-1999) or stagnated (1999-2009) when accounting for differential item functioning. We recommend that studies of the Flynn effect systematically investigate the possibility of differential item functioning to obtain unbiased ability estimates.

Keywords

Flynn effect; Negative Flynn effect; Wechsler scales; WAIS; Differential Item Functioning (DIF)

Highlights

- Items can change in their difficulty or discrimination over time.
- We tested the impact of these changes on Flynn effects in 3 versions of the WAIS.
- Over half the items showed significant differential item functioning.
- Differential item functioning biased the Flynn effect by about 3 IQ points per decade.
- The Flynn effect should be studied at the item level to obtain unbiased estimates.

1. Introduction

The Flynn effect refers to IQ changes over time in a population. First observed at the beginning of the 20th century (Rundquist, 1936), reported IQ changes over time have been overwhelmingly positive, with an average rate of about +3 IQ points per decade (Pietschnig & Voracek, 2013; Trahan et al., 2014). These gains seem to continue either at a similar rate (Trahan et al., 2014) or at a slowed rate in developed countries (Pietschnig & Voracek, 2013; Wongupparaj et al., 2015), although a few instances of a negative Flynn effect – IQ decreasing over time – have also been reported (Dutton & Lynn, 2015).

These IQ changes over time are often interpreted as long-term changes of *intelligence*, but this is not necessarily the case. By definition, a fluctuation of IQ is only a fluctuation of the total score on an intelligence test – and total scores on intelligence tests are only indirect reflections of intelligence. The score on an intelligence test is affected by many variables other than intellectual ability, such as cultural knowledge (e.g. Georgas et al., 2003; Kan et al., 2013) and test-taking strategies (Must & Must, 2013). If these other variables also change over time, they can lead to systematic overestimation or underestimation of scores for a sample at one point in time compared to another. In this case, the estimate of the Flynn effect will be biased: the actual change of intellectual ability may be less, or more, than the change occurring in the observed total score. In the current study, we focus on the possibility that estimates of the Flynn effect could be biased by changes of item difficulty over time (for a detailed discussion, see Gonthier et al., 2021).

1.1. Differential Item Functioning over Time

The literature has often raised the question of whether the Flynn effect reflects actual gains of intellectual ability, or just methodological artifacts (e.g. Kaufman, 2010; Rodgers, 1998; Weiss et al., 2016; Zhu & Tulsky, 1999). In his writings, James Flynn (e.g. 1998a,

2009) always steered clear of equating IQ changes with intelligence changes, taking up the useful analogy of Jensen (1994): that inferring changes of intellectual ability based on changes of test scores is akin to inferring differences of height based on differences in the length of people's shadows. Comparing the length of shadows collected at a particular point in time will yield accurate results, but comparing the length of shadows collected at different seasons, when the sun is lower or higher on the horizon, can yield biased estimates of height differences.

One factor that can particularly bias comparisons of total scores on intelligence tests over time is the change of item parameters: systematic changes, over time, in the difficulty or discriminating power of items (sometimes called *item drift*). As culture evolves over time, people approach the test with different cultural knowledge, making some items easier or more difficult. A useful example is given by Wicherts (2007; see also Wicherts et al., 2004), of people having higher success on a vocabulary test item requiring the definition of the word "terminate" after the release of the movie *Terminator,* and lower success on an item requiring the definition of the word "Kremlin" after the release of the movie *The Gremlins*. In these examples, scores change, but intellectual ability does not: changes of average performance are caused by changes in the attributes of items over time, due to a variable other than ability.

This situation, where differences of average performance between two samples are caused by differences in the items' difficulty or their capacity to discriminate between ability levels, rather than by differences of ability, is labeled Differential Item Functioning (DIF; for examples, see Ackerman, 1992; Martinková et al., 2017; Zumbo, 2007) – in this case DIF over time. DIF is often tested by examining difficulty and discrimination parameters at the item level based on Item Response Theory (IRT; e.g. Beaujean & Osterlind, 2008; Pietschnig et al., 2013). IRT allows not only for a test of differences of item parameters between samples, but also for a test of the impact of these differences on ability estimates. This makes

it possible to obtain estimates of ability differences between samples, independently of differences of item properties (as long as at least some items unbiased by DIF are available as a point of reference), a feature that could be particularly useful when testing for Flynn effects.

1.2. Impact of DIF over Time on the Flynn Effect

The phenomenon of DIF can change the difficulty of items over time, independently of any change of intellectual ability. In principle, this DIF over time can bias estimates of the Flynn effect (assuming that the Flynn effect reflects an actual change of ability); the result of the comparison between samples will depend on the direction in which item difficulty and sample ability change. If obsolescence leads items to become more difficult over time, this can partly or fully offset any long-term gains in ability in the population, leading to underestimation of the Flynn effect, or even to the ersatz finding of a negative Flynn effect (Gonthier et al., 2021). Conversely, items becoming easier over time could lead to overestimation of the Flynn effect.

Given the importance of this potential bias, there has been surprisingly little study of the role of DIF over time in intelligence tests, and its impact on estimates of the Flynn effect in particular. It has long been known that the items used in intelligence tests do indeed demonstrate systematic changes of difficulty over time, possibly affecting comparison between samples (Flieller, 1988; see also Brand et al., 1989). However, the extent of these changes and and their impact on estimates of the Flynn effect are stil unclear.

Some studies have confirmed that composite intelligence tests, such as the Wechsler scales, are not measurement invariant over time (Beaujean & Sheng, 2014; Wicherts et al., 2004), which means their measurement properties can indeed change over time. It has also been shown that this lack of measurement invariance can bias differences of latent means between samples collected at different points in time (Wicherts et al., 2004). However, these

studies conducted analyses only at the level of total scores on subtests, which makes it unclear how the lack of measurement invariance plays out at the level of items.

A handful of other studies have examined DIF using an IRT approach, but only in vocabulary and mathematics tests (Beaujean & Osterlind, 2008; Beaujean & Sheng, 2010; Flieller, 1988; Pietschnig et al., 2013). Most converged to the conclusion that there was significant item drift in these tests, with one study finding that DIF over time largely accounted for the Flynn effect (Beaujean & Osterlind, 2008). However, it is unknown to what extent this conclusion can be generalized to intelligence tests beyond the specific case of vocabulary and mathematics. One study (Shiu et al., 2013) investigated a panel of eight subtests, more diverse although still oriented towards verbal and numeric content (e.g. computation, information, sentence completion, synonyms), and found that over one third of all items demonstrated DIF. The direction and magnitude of DIF at the item level were not reported, but it had sufficient impact to severely bias estimates of the Flynn effect, at least for the information subtest: raw scores showed a negative Flynn effect for this subtest, whereas IRT-based ability estimates showed a positive Flynn effect.

We recently showed that a purported negative Flynn effect in France (Dutton & Lynn, 2015; see also Woodley of Menie & Dunkel, 2015) in the Wechsler scales was in fact driven by DIF over time, for some items in the subtests with high cultural load (Arithmetic, Comprehension, Information, Similarities, Vocabulary). Our results confirmed that DIF can indeed substantially bias Flynn effects, possibly contributing to the creation of ersatz negative Flynn effects due to outdated items becoming more difficult over time (Gonthier et al., 2021). To our knowledge, this was one of the only investigations of DIF using IRT in a test of general intelligence, in the context of Flynn effects. However, this study was only geared towards testing the possibility of a negative Flynn effect in France, and the generalizability of our conclusions to other contexts was limited by the small size of the sample ($N = 81$). A

systematic investigation of the contribution of DIF over time to Flynn effects in a general intelligence test is thus lacking. This is the focus of the present study.

1.3. Rationale for the Current Study

The overarching goal of the current study was to investigate the possibility that Flynn effects could be biased by DIF in general intelligence tests. This required answering two questions: 1) whether DIF over time is present in a test of a general intelligence, and to what extent; and 2) how unbiased estimates of the Flynn effect accounting for DIF, based on IRT ability estimates, compare to estimates of the Flynn effect computed from raw total scores, without correction for DIF.

Answering these two questions required a sample large enough to allow for stable IRT analysis; representative enough of the general population to allow for conclusions regarding the Flynn effect; and collected using a test of intellectual ability with enough different subtests to allow for general conclusions regarding the presence of DIF over time in intelligence tests. The only datasets matching these criteria in our country are the normative samples collected in the process of developing Wechsler scales.

A study of DIF also requires data collected with the same items at several successive points in time. There are three major ways to achieve this. The first solution is to have subjects complete the same test over years (e.g. Teasdale & Owen, 2008); but this is not the case for Wechsler scales, which are updated on a regular basis. The second solution is to have a small sample of subjects perform an older version of the test, and to compare their results with their performance on a newer version of the test in relation to normative samples (e.g. Flynn, 1984, 1998b); this is the solution we used in a prior study of DIF (Gonthier et al., 2021), but the resulting samples tend to be too small for large-scale IRT analyses. In the current study, we used a novel, third solution: taking advantage of the fact that some items

are re-used in successive versions of the same test, and testing DIF only for those items that overlap between successive versions[1].

We thus retrieved item-level datasets for three versions of the Wechsler Adult Intelligence Scale (WAIS): the WAIS-R dataset collected in 1989 (Wechsler, 1989), the WAIS-III dataset collected in 1999 (Wechsler, 2000), and the WAIS-IV dataset collected in 2009 (Wechsler, 2011). We identified the subset of items common to the WAIS-R and WAIS-III, and the subset of items common to the WAIS-III and WAIS-IV. We then treated these overlapping items as a single test, and we investigated whether these items demonstrated DIF, by comparing IRT item parameters between the 1989 and 1999 samples, and between the 1999 and 2009 samples. Lastly, we estimated the Flynn effect for the 1989-1999 and 1999-2009 periods based on the sum of scores on these items, and we compared these estimates of the Flynn effect with those obtain from IRT ability estimates accounting for DIF over time.

[1] In a sense, this method is symmetrical to the solution used by Flynn (1984): we use as a point of reference the common set of items that overlap between two versions of the test, instead of using a common set of subjects that perform two versions of the test.

2. Method

2.1 Datasets

The French publisher authorized access and use of the raw data for the normative samples of the WAIS-R (year 1989, *n* = 1000), WAIS-III (year 1999, *n* = 1104), and WAIS-IV (year 2009, *n* = 875). The three samples were approximately representative of the adult French population in terms of gender (WAIS-R: 50% male; WAIS-III: 45% male; WAIS-IV: 49% male), age groups (WAIS-R: 100 subjects in each of 10 groups in the 16-80 age range; WAIS-III: between 76 and 103 subjects in each of 12 groups in the 16-90 age range; WAIS-IV: between 67 and 87 subjects in each of 11 groups in the 16-90 age range), geographical regions (WAIS-R: between 136 and 271 subjects in each of 5 French territorial areas; WAIS-III: between 153 and 329 subjects in each of 5 French territorial areas; WAIS-IV: information unavailable but similar data collection methods), and socio-economic levels (approximately matching the composition of the general population, as assessed based on the categories of the French national institute of statistics, INSEE). All data were collected by psychologists purposefully trained by the publisher for WAIS data collection (each psychologist sent back protocols to the publisher after training to ensure that they complied with data collection instructions and that the test was scored correctly).

2.2 Subtest and Item Matching across Versions

Materials from the WAIS-R, WAIS-III and WAIS-IV were screened to identify items common to at least two test versions. Some subtests not scored as discrete items (e.g. Digit Symbol Coding) were discarded. To ensure that the distribution of scores was appropriate for DIF analyses, items with accuracy above 97.5% were excluded, as were items located before starting points, which were not completed by most subjects (e.g. Item 3 for a subtest starting at Item 4).

In most cases, items were strictly identical, or came with cosmetic changes (e.g. for the Picture Completion subtest, images of better quality in the WAIS-III than in the WAIS-R), but in 21 instances items were more substantially adapted from one version to the next. These 21 items were examined independently by the two authors to determine whether they could be considered logically equivalent. Eight of these items were considered logically equivalent by both authors, and were retained for analysis (these items are marked separately in the Results section); the others were discarded. The total number of items retained for analysis for each subtest is summarized in Table 1.

In some cases for the subtests Comprehension, Information, Similarities and Vocabulary, items were strictly identical, but the criteria used to score answers were altered from one version to the next. These changes were often minor: for example, one Vocabulary item of the WAIS-R had 27 scoring guidelines, of which 26 were kept constant for the WAIS-III, whereas the 27th was changed to allow the examiner to query one particular type of incomplete answer, giving the subject a chance to elaborate. In most cases, scoring criteria became more lenient (11 items), sometimes more stringent (3 items) or with a mix of more lenient and more stringent changes (4 items). All concerned items are marked separately in the Results section.

[Insert Table 1 approximately here]

2.3 Data Preprocessing

The data from the WAIS-R, WAIS-III and WAIS-IV were carefully preprocessed to ensure that they could be unbiasedly compared across the three samples. Subjects belonging to one of the clinical subsamples collected by the publisher were first excluded from the sample. The raw scores of all subjects were then retrieved for all items in all subtests. Data entry errors were corrected in all datasets. Missing data for certain items, due to the subject

reaching the discontinue criterion in a subtest, were recoded as 0 for the three versions[2]. For the three subtests including items scored as a function of response time (Arithmetic, Block Design and Object Assembly), responses were re-scored for those items where time credit differed across versions. Responses were also re-scored for the Picture arrangement subtest, where different criteria for partly correct responses were used in the WAIS-R and WAIS-III. To ensure stability of the estimated parameters, we also recoded items with more than two possible scores where a given score was obtained by fewer than 25 subjects in a given test version, by collapsing the response category with insufficient data with the immediately inferior response (for example, in the case of an item scored 0, 1 or 2, when only five subjects scored 1, the item was recoded as 0 or 1 and these five subjects were assigned a score of 0).

2.4 Data Analysis

Differential item functioning was tested by comparing the WAIS-R and WAIS-III samples on one hand, and the WAIS-III and WAIS-IV samples on the other hand: there were too few identical items between WAIS-R and WAIS-IV to allow for meaningful comparison (n = 15). Analyses were performed with the method of iterative logistic ordinal regression using IRT (Choi et al., 2011; Crane et al., 2006), as implemented with the package *lordif* (Choi, 2016; see also Choi et al., 2011) for R (R Core Team, 2022).

Logistic regression can be used to test how the score on a given item varies as a function of both a subject's ability, and the group to which they belong; this is a classic and robust approach to DIF (Swaminathan & Rogers, 1990). Logistic *ordinal* regression is an

[2] This was done to maximize the amount of data available for ability estimation, with the side effect that the more difficult items were scored as failed despite subjects not completing them due to failing prior items, potentially biasing item parameter estimates. However, replacement by zero seems to have limited effect on Type I error rates when the data are not missing at random (Banks, 2015), and the current results were relatively robust to this analytic choice: when coding missing data as "NA" instead of zero, 32 items instead of 34 had significant uniform DIF for the comparison between WAIS-R and WAIS-III, and 25 items instead of 29 had significant uniform DIF for the comparison between WAIS-III and WAIS-IV.

extension of logistic regression to the case of dependent variables with more than two outcomes. This allows for the analysis of a mixture of items with two or more than two possible scores, which is particularly useful in the case of the WAIS.

For each item, three models are compared: Model 1 predicts item score based only on ability; Model 2 predicts item score based on both ability and group; Model 3 predicts item score based on ability, group, and the interaction between the two. If Model 2 fits better than Model 1, the item has uniform DIF (scores on the item depend on the subject's group, above and beyond their ability); if Model 3 fits better than Model 2, the item has non-uniform DIF (the relation between level of ability and scores on the item depends on group). Note that an item can have only uniform DIF (the intercept is higher in one group, indicating lower difficulty, but the relation between ability and performance is the same in both groups), only non-uniform DIF (the slope for the effect of ability on performance is lower in one group, but average difficulty is the same), or both. In the current study, we estimated the difference between models using Nagelkerke's pseudo-R^2 measure of explained variation[3] (Nagelkerke, 1991; among available alternatives, the values of this index tend to be closest to the equivalent R^2 in a multiple regression, e.g. Veall & Zimmermann, 1990).

With the method of *iterative* logistic ordinal regression (Choi et al., 2011; Crane et al., 2006), the first step is to estimate ability by fitting an IRT model to all items, assuming that no DIF is present. The corresponding theta parameter estimates are retrieved and serve to index ability. In the second step, a logistic ordinal regression is used to identify items with substantial DIF, as described above. The IRT model is then fitted again, with separate parameters for the two versions for all items identified with DIF, so as to obtain a more precise ability estimate. This procedure is run iteratively until all items with DIF are

[3] Tables 3 to 6 only report the comparison between Model 1 and Model 2 for uniform DIF, and the comparison between Model 2 and Model 3 for non-uniform DIF. A two-degrees of freedom comparison between Model 1 and Model 3 is also possible to test for overall DIF, but it is not reported here for simplicity.

identified. In the current study, IRT estimation used the graded response model (Samejima, 1969) with default parameters. Items were flagged with substantial DIF in the logistic ordinal regression if the difference of R^2 between Model 1 and Models 2 or 3 was at least 0.01. (Other possible criteria, such as a R^2 of 0.02 or a significant chi-square test, led to similar conclusions.)

After performing this iterative procedure to obtain stable ability estimates, pseudo-R^2 were computed for the difference between Model 1, Model 2 and Model 3 for each item to quantify the effect size for DIF. A significance test was also performed by conducting Monte Carlo simulations under the null hypothesis with 5000 replications, so as to obtain approximate *p*-values for these R^2. The significance threshold for DIF was set at alpha = .001 to correct for multiple comparisons across all items.

The last step was to estimate the extent of the Flynn effect, with and without taking DIF into account. To this end, approximate IQ scores were computed for each subject, based on raw item responses (the scores on all items were normalized on a scale from 0 to 1, summed together, and converted to the standard IQ scale), and based on IRT ability estimates corrected by DIF (computed as theta ability estimates, with separate parameters for items with DIF, then converted to the standard IQ scale). This allowed for comparison of the raw estimate of the Flynn effect that would have been obtained based on simply counting correct answers[4], to the more refined estimate obtained from IRT allowing for DIF, in line with prior literature (Beaujean & Osterlind, 2008; Beaujean & Sheng, 2010; Pietschnig et al., 2013).

[4] An alternative solution would have been to use the IRT ability estimates in a model not allowing for DIF (i.e. with all item parameters constrained to be equal across test versions). This alternative led to conclusions similar to using the sum of raw item responses, with a Flynn effect estimated to +1.93 IQ points for the 1989-1999 comparison and to -2.37 IQ points for the 1999-2009 comparison.

3. Results

The results of DIF analyses are summarized in Table 2. Details of the statistical tests for all subtests are presented in Tables 3-6; details of IRT item parameters are presented as a Supplemental Material in Tables A1-A4. In total, over half the items demonstrated DIF over time. This was true both between the 1989 WAIS-R sample and the 1999 WAIS-III sample, and between the 1999 WAIS-III sample and the 2009 WAIS-IV sample.

[Insert Table 2 approximately here]

The majority of observed DIF was uniform DIF (a difference of intercept: items being significantly more difficult in one sample than another, for the same level of intellectual ability), which occurred for over half the items. By contrast, there were fewer instances of non-uniform DIF (a difference of slope: items being significantly more dependent on ability for one sample than another). In total, about one third of items had significant non-uniform DIF, but half of these came from just the Information and Vocabulary items in the WAIS-R and WAIS-III samples.

Overall, there were more instances of DIF for subtests with a high cultural load (Georgas et al., 2003; Kan et al., 2013): over half the items had uniform DIF in all of Arithmetic, Comprehension, Information, Similarities, Vocabulary and Picture completion, whereas there were only 38% of items with uniform DIF in Matrix reasoning, and almost none in Block design. Surprisingly, there was significant DIF for all analyzed items of the Backward digit span.

The detailed results of the comparison between the 1989 WAIS-R and 1999 WAIS-III samples are displayed in Table 3 for verbal subtests and in Table 4 for visuo-spatial subtests. Note that the results are split between these two tables for legibility, but that all items were analyzed concurrently to obtain the ability estimates. It is clear from the results that the vast

majority of items belonging to verbal subtests had uniform DIF, non-uniform DIF, or both. It is also clear that the magnitude of DIF, expressed in terms of pseudo-R^2, was relatively small (although pseudo-R^2 from logistic regressions do not translate directly into a percentage of explained variation, and are typically lower than those from linear regressions): for items strictly identical across versions, effect sizes ranged from $R^2 = .01$ to .04. Uniform DIF was mostly in the direction of items being more difficult for the 1999 WAIS-III sample than for the 1989 WAIS-R sample for an equal level ability (24 out of 34 items or 71%).

[Insert Table 3 approximately here]

[Insert Table 4 approximately here]

The detailed results of the comparison between the 1999 WAIS-III and 2009 WAIS-IV samples are displayed in Table 5 for verbal subtests and in Table 6 for visuo-spatial subtests. In this case too, the majority of items in the verbal subtests demonstrated uniform DIF, non-uniform DIF or both. For most items, DIF effect sizes were in the $R^2 = .01$ to .07 range. There were two exceptions for the Picture completion subtest, at .11 and .50: upon closer inspection, this was explained by the redrawing of the two corresponding items in the WAIS-IV, with the enhanced level of detail in the pictures making the missing features much less perceptually obvious. Uniform DIF was mostly in the direction of items being more difficult for the 2009 WAIS-IV sample than for the 1999 WAIS-III sample for an equal level ability (21 out of 29 items or 72%).

[Insert Table 5 approximately here]

[Insert Table 6 approximately here]

Overall and across the three WAIS versions, the majority of analyzed items became more difficult over time for the Arithmetic, Digit Span, Picture Completion, and Vocabulary

subtests. Picture Arrangement and Object Assembly had few items, but also demonstrated DIF in the direction of being more difficult. Comprehension, Information, and Similarities had a mix of items becoming more difficult and less difficult. There was little DIF for Block Design and Matrix Reasoning, and all instances of DIF were for items becoming easier.

For items whose scoring criteria changed across versions, DIF was in the direction opposite to scoring changes in four cases (e.g. scoring became more lenient whereas DIF indicated that the item became comparatively harder); in these cases, DIF was probably underestimated. In five more cases, DIF was non-significant despite a change of scoring criteria; in these cases, scoring changes potentially masked the presence of DIF. In four other cases, DIF was in the same direction as scoring changes; in these cases, scoring changes potentially explained the presence of DIF. The last five cases were ambiguous due to the presence of non-uniform DIF or due to scoring criteria becoming both more lenient and more stringent. In sum, changes of scoring potentially explained 9 out of 76 instances of DIF, and potentially led to underestimating DIF in 9 other cases.

The final step of the analysis was to compare the Flynn effect estimated based on the sum of raw item scores, and based on theta ability estimates corrected for the presence of DIF. For the comparison between 1989 WAIS-R and 1999 WAIS-III, based on raw scores the estimated Flynn effect was +1.03 IQ points (IQ = 99.46 for the 1989 WAIS-R sample and IQ = 100.49 for the 1999 WAIS-III sample); based on theta ability estimates corrected for DIF, the estimated Flynn effect was +3.87 IQ points (IQ = 97.97 for the 1989 WAIS-R sample and IQ = 101.83 for the 1999 WAIS-III sample), closer to the expected rate (Pietschnig & Voracek, 2015; Trahan et al., 2014). In other words, raw item scores underestimated the Flynn effect by 2.84 IQ points.

For the comparison between 1999 WAIS-III and 2009 WAIS-IV, based on raw scores the estimated Flynn effect was -3.62 IQ points (IQ = 101.60 for the 1999 WAIS-III sample

and IQ = 97.98 for the 2009 WAIS-IV sample), suggesting a negative Flynn effect (Dutton & Lynn, 2015); based on theta ability estimates corrected for DIF, the estimated Flynn effect was +0.02 IQ points (IQ = 99.99 for the 1999 WAIS-III sample and IQ = 100.01 for the 2009 WAIS-IV sample), consistent with a slowing of the Flynn effect (Pietschnig & Voracek, 2015) but not with a negative Flynn effect. In other words, raw item scores underestimated the Flynn effect by 3.64 IQ points.

4. Discussion

Our analysis of DIF in Wechsler subtests led to six major conclusions. 1) There was substantial evidence of DIF over time, with over half of all items demonstrating significant differential functioning across the 1989 WAIS-R, 1999 WAIS-III and 2009 WAIS-IV samples, despite a conservative significance threshold set at $p = .001$. 2) DIF was more prevalent in some subtests than others; Block Design and Matrix Reasoning tests were least affected, although not immune to DIF. 3) Observed instances of DIF were mostly uniform DIF, indicating higher difficulty for one sample than another; non-uniform DIF, indicating higher discriminating ability in one sample than another, was less prevalent. 4) Uniform DIF was mostly in the direction of items becoming more difficult over time for the same level of ability; this was true for a little over two third of the items demonstrating DIF. 5) The effect size for DIF was generally low for strictly identical items, although a few items had DIF up to $R^2 = .07$. 6) Despite the DIF effect size being relatively low for each separate item, its cumulative impact across the whole test led to substantial bias in estimates of the Flynn effect: the progression of IQ scores was underestimated by 2.84 IQ points between the 1989 and 1999 samples and by 3.64 IQ points between the 1999 and 2009 samples.

Overall, these findings converge with prior literature in showing that there can be substantial variations over time in the difficulty of tests of intellectual ability (Beaujean &

Osterlind, 2008; Flieller, 1988; Pietschnig et al., 2013; Shiu et al., 2013; Wicherts, 2007; Wicherts et al., 2004), and that these variations of difficulty can directly bias estimates of the Flynn effect, substantially affecting the conclusions drawn about long-term fluctuations of intelligence (Beaujean & Osterlind, 2008; Beaujean & Sheng, 2014; Shiu et al., 2013; Wicherts et al., 2004) and making IRT-based estimates of ability inherently preferable (Beaujean & Osterlind, 2008; Beaujean & Sheng, 2010; Pietschnig et al., 2013).

4.1. Impact of DIF on the Flynn Effect

The misestimation of the Flynn effect introduced by DIF over time was sufficient to substantially affect the conclusions that could be drawn based on the current dataset. For the 1989-1999 period, the raw difference in scores suggested a minimal Flynn effect at +1.03 IQ points in a decade, compatible with a slowing in the Flynn effect (Pietschnig & Voracek, 2015), whereas the actual figure was +3.87 IQ points in a decade, close to the average value of about three points per decade for the effect (Trahan et al., 2014). For the 1999-2009 period, the raw difference in scores suggested a negative Flynn effect at -3.62 IQ points, very close to the value of -3.8 points claimed by Dutton and Lynn for France (2015); whereas the actual figure was a positive Flynn effect of +0.02 points, consistent with a recent slowing down or interruption of the Flynn effect in recent times for developed countries (Pietschnig & Voracek, 2015; Wongupparaj et al., 2015), but not with an intelligence decline.

These findings generally converge with our prior work with the French WAIS (Gonthier et al., 2021): they confirm, in a much larger sample representative of the French population, that there is indeed no negative Flynn effect in France (although there most likely is a decline in the magnitude of the Flynn effect, in line with the literature), with the observation of raw score declines in some subtests being attributable to drifts in item difficulty over time.

Beyond the specific case of France, this highlights the necessity of systematically investigating the possibility of DIF in all assessments of the Flynn effect. Given that the bias introduced by DIF was around 3 IQ points per decade in the current study, in the same range as the Flynn effect itself, Flynn effect studies are certainly at risk of DIF hiding the true changes of average intellectual ability over time. Assuming that the Flynn effect is slowing down in developed countries (Pietschnig & Voracek, 2015; Wongupparaj et al., 2015), the presence of DIF over time in the direction of items becoming more difficult could be enough to offset small gains in intellectual ability, and spuriously create the recent findings of negative Flynn effects (Dutton et al., 2016).

More generally, these findings complement prior studies investigating tests with more restricted content (Beaujean & Osterlind, 2008; Shiu et al., 2013), or examining the WAIS and other test batteries at the subtest rather than item level (Beaujean & Sheng, 2014; Wicherts et al., 2004), which also converged to the conclusion that non-constant difficulty of tests or items can bias Flynn effect estimates. By contrast, a single study concluded that IRT-based estimates of the Flynn effect were in the same order of magnitude as estimates based on sum scores, although somewhat smaller (Pietschnig et al. 2013). The one study finding no DIF over time (Beaujean & Sheng, 2010) also found that IRT-based estimates of the Flynn effect were substantially higher than estimates based on sum scores, comparable to our own results, and further encouraging the use of IRT in future studies on the Flynn effect.

4.2. Expected DIF for Different Items

At the item level, our results showed that most of the biased items became more difficult over time, and that DIF was generally of low magnitude. These two findings are not expected to generalize to all studies of DIF over time in intelligence tests, and are probably due to the method used here: we analyzed only those items common to consecutive versions of the WAIS. By definition, this means that the items included in the current study were

created for an older sample, then screened by the test developer to ensure that they remained current enough to be re-used for a newer version. This has two implications. First, items with high expected DIF over time were presumably not included by the test developer in the next version, reducing the magnitude of observed DIF. In other words, the current results probably underestimate the possible magnitude of DIF over time when tests are not updated: in prior studies with the French WAIS (Dutton & Lynn, 2015; Gonthier et al., 2021), subjects in 2009 and 2019 were asked to perform *all* items from the 1999 version of the test, potentially leading to greater DIF. Second, items were presumably more likely to become outdated, and thus to become more difficult to solve as their contents become less well-known over time.

It is not always explicit when inspecting the items why their answers should become less well-known over time, although some hypotheses can be made (see also Gonthier et al., 2021). Most items became more difficult for the Information subtest, which is largely based on knowledge of famous people and works of art from the XXth century; it is expected that this knowledge will fade from public knowledge over time. Items becoming more difficult for the Arithmetic subtest may be related to the continuous decline in math knowledge in France (OECD, 2019). For the Vocabulary subtest, this may be related to words falling out of use in the language. The Picture Completion subtest primarily depicts objects common in the XXth century and rural scenes, which can be expected to be less familiar to modern test-takers. For other subtests, such as Digit Span Backward, Block Design or Matrix Reasoning, there is no obvious explanation for the presence of DIF.

Prior studies about DIF over time in tests of intellectual ability have agreed neither on the extent, nor on the direction of DIF. The current data showed DIF in about half of items, compared to about one sixth (Pietschnig et al., 2013), one third (Gonthier et al., 2021; Shiu et al., 2013), half (Beaujean & Osterlind, 2008), or two thirds (Flieller, 1988) of items. We found that most items became more difficult over time, leading to underestimates of the

Flynn effect; similar results were found in some studies (Gonthier et al., 2021; Shiu et al., 2013), but other studies found that items became easier or that variations of difficulty led to overestimates of the Flynn effect (Beaujean & Osterlind, 2008; Pietschnig et al., 2013); yet other studies found a mix of items becoming more and less difficult, and variations of difficulty leading to both underestimates and overestimates (Beaujean & Sheng, 2014; Flieller, 1988; Wicherts et al., 2004). In practice, it is expected that the extent and direction of DIF will differ based on the type of items and the type of knowledge they require. As a result, no general conclusion can be made, except to stress that the presence of DIF on at least some items is very likely and can introduce unpredictable bias.

4.3. Expected DIF for Different Subtests

We found that DIF was substantially less prevalent in Block design and Matrix reasoning than in other subtests, which suggests that drifts of difficulty over time are related to the cultural load of a subtest (Georgas et al., 2003; Kan et al., 2013). In other words, subtests which involve cultural knowledge to a larger extent, especially declarative knowledge, are liable to demonstrate more impact of DIF over time (see Gonthier et al., 2021). In the Wechsler scales, this prominently includes the Arithmetic, Comprehension, Information, Similarities and Vocabulary subtests, which all require subjects to answer questions based on knowledge acquired more or less explicitly (vocabulary words, general knowledge, social rules, etc) in a way highly specific to a given cultural context. The implication is that studies of the Flynn effect based on tests that make less use of this type of declarative knowledge, such as matrix reasoning tasks, will tend to yield Flynn effect estimates less biased by DIF.

However, close inspection of items affected by DIF in the current study shows that items from all subtests could be affected – including Block Design and Matrix Reasoning – which suggests that this is a very general phenomenon. This also constitutes a reminder that

no test is really exempt from cultural influences. Even visuo-spatial tests such as matrix tasks and constructive tasks require culturally acquired procedural knowledge, such as reading the item in a certain direction, paying attention to exact numerosity, being familiar with certain shapes, and being used to playing with wooden blocks (for an extensive review, see Gonthier, 2022; see also Greenfield, 1997). Although these pieces of knowledge are probably less variable over time in a given culture than knowledge of trivia or vocabulary use, the current results suggest that long-term trends could occur as well. Moreover, certain visuo-spatial tests make heavy use of cultural concepts: this is the case for Picture Completion in the WAIS-III (which requires identifying missing features in pictures of scenes or objects expected to be familiar to the subject), where over two thirds of all items had uniform DIF. In short, the results confirm that visuo-spatial tests should also be screened for DIF over time.

The finding that DIF was generally less prevalent for visuo-spatial subtests than for tests making heavy use of declarative knowledge has one interesting implication for prior studies estimating the Flynn effect. It has repeatedly been found that the Flynn effect is larger for tests of fluid intelligence; by contrast, tests of crystallized intelligence show smaller gains and are more likely to demonstrate an interruption of the Flynn effect (Pietschnig & Voracek, 2015; for an illustration, see Flynn, 2009). Given that fluid intelligence is usually measured with visuo-spatial subtests such as matrices, and crystallized intelligence is usually measured with tests of declarative knowledge such as vocabulary and arithmetic, estimates of the Flynn effect can be expected to be more biased by DIF for crystallized intelligence. Furthermore, we found that DIF over time is mostly in the direction of items becoming more difficult, leading to an underestimate of the Flynn effect (see also Gonthier et al., 2021); if this finding holds more generally in other datasets, this may partly explain why crystallized intelligence shows smaller gains than fluid intelligence: the Flynn effect may be partly compensated by increasing difficulty at the item level.

4.4. Methods of Testing for Flynn Effects and DIF

The new method proposed here to test DIF over time, based on items overlapping across successive versions of the same test, allowed us to gain insight into the change of item parameters across two decades in representative samples of the general population. This was particularly helpful in our country where large-scale intelligence testing is rarely performed, and where there was no other way to test this hypothesis. It is also one of the very few possibilities available to test intelligence trends over time on the basis of existing data. We believe this makes it a useful addition to the toolbox of intelligence researchers.

However, the method of using overlapping items as described in the current study is far from perfect. Its major issue is that it cannot control for changes in the context in which an item is performed in the test (see Zwick, 1991). This includes subtle changes in the way in which instructions are worded, or in which responses are scored; changes in the order of subtests within the test (which could affect cognitive fatigue or disengagement); and most problematically, the position of items within a subtest. In the current datasets, some items were identical but performed at different points in successive versions of the test, which can bias the results in various ways: subjects completing an item at a later point in the test have received more training, but have a higher likelihood of not completing the item at all due to reaching the discontinue criterion on prior items. In this study, we ensured that DIF was present even for items completed at the beginning of a subtest (see Tables 3-6), and even when scoring missing values as "NA", which partly mitigates the latter issue. We recommend that the same precaution be taken in future studies using the same method, along with careful consideration of small methodological changes, including changes of scoring (see Section 2.3 Data Preprocessing).

While context effects are an actual concern for our method, there are not many alternatives to test Flynn effects and DIF over time when different tests are performed at

different timepoints: the only other existing method is to have a group of subjects complete both the older and newer version of the same test to serve as a point of comparison (e.g. Dutton & Lynn, 2015; Flynn, 1984, 1998b). This method has its own problems, primarily related to small and unrepresentative samples (see Gonthier et al., 2021). In this light, we believe the method of overlapping items described here to be a helpful complement – and one which can be particularly useful in other datasets using tests that experience less changes than successive versions of the WAIS. The two methods of using a common set of subjects or a common set of items can even be used in parallel to confirm each other's conclusions (just like the current results appear to confirm the conclusions of Gonthier et al., 2021).

A possible extension of our method would be to use overlapping items as anchors for test linking (for an introduction to this topic, see Kolen & Brennan, 2014; for an example, see Shiu et al., 2013). The idea of test linking is to use items common to two versions of a test as a point of reference to place the IRT parameters of other items on the same scale for the two versions. This makes it possible to use data from *all* items to obtain ability estimates that are directly comparable between the two versions; by contrast, our study used just the overlapping items themselves. Test linking is a powerful method, but is only appropriate when major precautions are met: there must be enough overlapping items without DIF to serve as anchors (Kolen & Brennan, 2014, recommend that they represent 20% of all items), and these items should be spread evenly across difficulty levels and test content, criteria which were not met in the current dataset. It is also critical that overlapping items serving as anchors be presented in similar contexts (Kolen & Brennan, 2014; Zwick, 1991), which as discussed above, is not the case in consecutive versions of the Wechsler scales, making test linking generally unsuitable in this case. However, the method can be useful with intelligence tests including more similar or even identical content (Shiu et al., 2013).

Critically, methodological limitations regarding context effects are inherent to the particular case of testing for DIF over time based on items overlapping across successive versions of the same test. When on the other hand the same test is completed by all subjects across years, an unbiased test of DIF over time can be easily achieved. This is the case, for example, for large-scale testing of military conscripts using Borge Priene's Prove in Denmark (Teasdale & Owen, 2008) or the Peruskoe test in Finland (Dutton & Lynn, 2013), whose content does not change. Both tests have been used to claim negative Flynn effects (Dutton & Lynn, 2013; Dutton et al., 2016; Teasdale & Owen, 2008), but both tests include content with a substantial cultural load (e.g. verbal analogies and word knowledge), which makes them particularly exposed to DIF over time. These are just two examples. Given the current results and prior research (Pietschnig et al., 2013; Wicherts et al., 2004), we argue that all datasets used to investigate Flynn effects should be systematically screened for DIF over time.

Acknowledgements

The authors thank Pearson and the ECPA (*les Editions du Centre de Psychologie Appliquée*) for authorizing access to the WAIS-R, 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

Banks, K. (2015). An Introduction to missing data in the context of Differential Item Functioning. *Practical Assessment, Research, and Evaluation*, *20*(12). https://doi.org/10.7275/fpg0-5079

Beaujean, 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

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

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

Choi, S. W., Gibbons, L. E., & Crane, P. K. (2011). lordif: An R Package for Detecting Differential Item Functioning Using Iterative Hybrid Ordinal Logistic Regression/Item Response Theory and Monte Carlo Simulations. *Journal of statistical software*, *39*(8), 1–30. https://doi.org/10.18637/jss.v039.i08

Crane, P. K., Gibbons, L. E., Jolley, L., & van Belle, G. (2006). Differential item functioning analysis with ordinal logistic regression techniques. DIFdetect and difwithpar. *Medical Care*, *44*(11 Suppl 3), S115 – 123. https://doi.org/10.1097/01.mlr.0000245183.28384.ed

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

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. https://doi-org.distant.bu.univ-rennes2.fr/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. https://doi.org/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. https://doi.org/10.2466/pms.1998.86.3c.1231.

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

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. https://doi.org/10.1016/B978-012280055-9/50021-7

Gonthier, C. (2022). Cross-cultural differences in visuo-spatial processing and the culture-fairness of visuo-spatial intelligence tests: An integrative review and a model for matrices tasks. *Cognitive Research: Principles and Implications*. https://doi.org/10.1186/s41235-021-00350-w

Gonthier, C., Grégoire, J., & Besançon, M. (2021). No negative Flynn effect in France: Why variations of intelligence should not be assessed using tests based on cultural knowledge. *Intelligence*, *84*. https://doi.org/10.1016/j.intell.2020.101512

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

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. https://doi.org/10.1177/0956797613493292

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. https://doi.org/10.1177/0734282910373346

Kolen, M. J. & Brennan, R. L. (2014). *Test equating, scaling, and linking (Third Edition)*. Springer. https://doi.org/10.1007/978-1-4939-0317-7

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. https://doi.org/10.1187/cbe.16-10-0307

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

Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, *78*, 691-693.

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

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. https://doi.org/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. https://doi.org/10.1177/1745691615577701

R Core Team (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

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

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

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, *34*(4, Pt. 2), 100.

Shiu, W., Beaujean, A. A., Must, O., te Nijenhuis, J., & Must, A. (2013). An item-level examination of the Flynn effect on the National Intelligence Test in Estonia. *Intelligence*, *41*(6), 770–779. https://doi.org/10.1016/j.intell.2013.05.007

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*(4), 361–370.

Teasdale, T. W., & Owen, D. R. (2008). Secular declines in cognitive test scores: A reversal of the Flynn Effect. *Intelligence*, *36*(2), 121–126. https://doi.org/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. https://doi.org/10.1037/a0037173

Veall, M., & Zimmermann, K. (1990). *Evaluating pseudo-R2's for binary probit models*. (CentER DiscussionPaper, Vol. 1990-57). Retrieved from: https://research.tilburguniversity.edu/files/1149062/MRVKFZ5620446.pdf

Wechsler, D. (1989). *Manuel de l'Echelle d'Intelligence de Wechsler Pour Adultes, forme révisée* [Manual for the Wechsler Adult Intelligence Scale – Revised Edition]. ECPA.

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. https://doi.org/10.1177/0734282915621222

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. https://doi.org/10.1016/j.intell.2004.07.002

Wongupparaj, P., Kumari, V., & Morris, R. G. (2015). A Cross-Temporal Meta-Analysis of Raven's Progressive Matrices: Age groups and developing versus developed countries. *Intelligence*, *49*, 1-9. https://doi.org/10.1016/j.intell.2014.11.008

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. https://doi.org/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. https://doi.org/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. https://doi.org/10.1080/15434300701375832

Zwick, R. (1991). Effects of item order and context on estimation of NAEP reading proficiency. *Educational Measurement: Issues and Practice*, *10*(3), 10–16. https://doi.org/10.1111/j.1745-3992.1991.tb00198.x

Table 1

Number of items retained for analysis in each subtest

Subtest	Analyzable items common to WAIS-R and WAIS-III	Analyzable items common to WAIS-III and WAIS-IV
Arithmetic	5	1
Block Design	9	4
Comprehension	3	4
Digit Span Forward	NA	5
Digit Span Backward	NA	5
Information	9	4
Matrix Reasoning	NA	8
Object Assembly	3	NA
Picture Arrangement	3	NA
Picture Completion	10	12
Similarities	5	5
Vocabulary	12	4
Total	*59*	*52*

Note. NA indicates that the subtest was not included in one version or that raw item data were not available.

Table 2

Number of items with uniform and non-uniform DIF per subtest

Subtest	WAIS-R and WAIS-III		WAIS-III and WAIS-IV	
	Uniform DIF	Non-uniform DIF	Uniform DIF	Non-uniform DIF
Arithmetic	5/5 (100%)	0/5 (0%)	0/1 (0%)	0/1 (0%)
Block Design	0/9 (0%)	3/9 (33%)	1/4 (25%)	0/4 (0%)
Comprehension	2/3 (67%)	1/3 (33%)	3/4 (75%)	0/4 (0%)
Digit Span Forward	-	-	1/5 (20%)	1/5 (20%)
Digit Span Backward	-	-	5/5 (100%)	0/5 (0%)
Information	4/9 (44%)	7/9 (78%)	3/4 (75%)	0/4 (0%)
Matrix Reasoning	-	-	3/8 (38%)	0/8 (0%)
Object Assembly	1/3 (33%)	2/3 (67%)	-	-
Picture Arrangement	3/3 (100%)	0/3 (0%)	-	-
Picture Completion	7/10 (70%)	4/10 (40%)	8/12 (67%)	3/12 (25%)
Similarities	3/5 (60%)	3/5 (60%)	3/5 (60%)	2/5 (40%)
Vocabulary	9/12 (75%)	10/12 (83%)	4/4 (100%)	0/4 (0%)
Total	*34/59 (58%)*	*30/59 (51%)*	*29/52 (56%)*	*6/52 (12%)*

Table 3

Comparison between WAIS-R and WAIS-III for verbal subtests

Item ID		Item	Uniform DIF			Non-uniform DIF		
WAIS-R	WAIS-III	differences	R^2	*p*-value	Direction	R^2	*p*-value	Direction
ARI-09	ARI-10		**.01**	**<.001**	**harder**	.00	.108	
ARI-10	ARI-14		**.02**	**<.001**	**harder**	.00	.458	
ARI-12	ARI-13		**.01**	**<.001**	**harder**	.00	.246	
ARI-13	ARI-18	!	**.06**	**<.001**	**harder**	.00	.088	
ARI-14	ARI-20	!	**.04**	**<.001**	**harder**	.00	.156	
COM-03	COM-04		**.01**	**<.001**	**easier**	.00	.006	
COM-06	COM-08		.00	.447		.00	.068	
COM-12	COM-12		**.02**	**<.001**	**harder**	**.01**	**<.001**	**less disc.**
INF-07	INF-07		.00	.801		.00	.141	
INF-13	INF-19		.00	.580		**.01**	**<.001**	**less disc.**
INF-14	INF-06		.01	.002		.00	.174	
INF-15	INF-12		**.01**	**<.001**	**easier**	**.01**	**<.001**	**less disc.**
INF-16	INF-18		.00	.709		**.01**	**.001**	**less disc.**
INF-22	INF-10	s-	**.05**	**<.001**	**easier**	**.02**	**<.001**	**less disc.**
INF-23	INF-15	s-	.00	.752		**.03**	**<.001**	**less disc.**
INF-24	INF-22		**.01**	**<.001**	**easier**	**.01**	**<.001**	**less disc.**
INF-26	INF-24		**.01**	**.001**	**harder**	**.03**	**<.001**	**less disc.**
SIM-04	SIM-06		.00	.378		.00	.084	
SIM-06	SIM-08		**.04**	**<.001**	**harder**	**.02**	**<.001**	**less disc.**
SIM-07	SIM-07		**.01**	**<.001**	**easier**	.00	.495	
SIM-11	SIM-12		**.04**	**<.001**	**harder**	**.03**	**<.001**	**less disc.**
SIM-13	SIM-13		.00	.002		**.02**	**<.001**	**less disc.**
VOC-05	VOC-04		**.01**	**.001**	**easier**	.00	.744	
VOC-09	VOC-06		**.01**	**<.001**	**easier**	**.01**	**<.001**	**less disc.**
VOC-13	VOC-07	s#	.00	.768		**.01**	**<.001**	**less disc.**
VOC-16	VOC-12		**.01**	**<.001**	**harder**	**.01**	**<.001**	**less disc.**
VOC-21	VOC-29		**.02**	**<.001**	**harder**	**.02**	**<.001**	**less disc.**
VOC-24	VOC-15		**.01**	**<.001**	**harder**	**.01**	**<.001**	**less disc.**
VOC-27	VOC-21	s+	**.01**	**<.001**	**harder**	**.01**	**<.001**	**less disc.**
VOC-28	VOC-23	s-	**.01**	**<.001**	**harder**	**.03**	**<.001**	**less disc.**
VOC-30	VOC-27	s+	**.02**	**<.001**	**harder**	**.02**	**<.001**	**less disc.**
VOC-32	VOC-24		.00	.239		**.02**	**<.001**	**less disc.**
VOC-33	VOC-32		.00	.714		**.02**	**<.001**	**less disc.**
VOC-35	VOC-33		**.02**	**<.001**	**harder**	.00	.340	

Note. These items were analyzed along with those in Table 4. ARI =Arithmetic, COM = Comprehension, INF = Information, SIM = Similarities, VOC = Vocabulary. Item differences are marked ! for items not strictly identical but logically equivalent, or s+, s-, and s# for identical items with different scoring criteria in the more recent version (respectively more stringent criteria, more lenient criteria, and both more stringent and more lenient criteria). R^2 is the Nagelkerke pseudo-R^2 from the logistic ordinal regression, *p* is the corresponding *p*-value based on Monte-Carlo simulations. Comparisons yielding significant DIF are in boldface.

Table 4

Comparison between WAIS-R and WAIS-III for visuo-spatial subtests

Item ID		Item	Uniform DIF			Non-uniform DIF		
WAIS-R	WAIS-III	differences	R^2	*p*-value	Direction	R^2	*p*-value	Direction
ARR-06	ARR-06		**.02**	**<.001**	**harder**	.00	.006	
ARR-09	ARR-08		**.03**	**<.001**	**harder**	.00	.028	
ARR-10	ARR-07		**.03**	**<.001**	**easier**	.00	.018	
BD-01	BD-05		.01	.006		.00	.247	
BD-02	BD-07		.01	.016		.00	.094	
BD-03	BD-06		.02	.004		.00	.443	
BD-04	BD-08		.00	.098		.00	.690	
BD-05	BD-09		.00	.302		.00	.004	
BD-06	BD-10		.00	.152		.00	.203	
BD-07	BD-11	!	.00	.051		**.01**	**<.001**	**more disc.**
BD-08	BD-12		.00	.684		**.00**	**<.001**	**more disc.**
BD-09	BD-13		.00	.094		**.01**	**<.001**	**more disc.**
OBA-01	OBA-01		.00	.054		.00	.138	
OBA-02	OBA-02		.00	.275		**.00**	**<.001**	**more disc.**
OBA-04	OBA-03		**.02**	**<.001**	**harder**	**.01**	**.001**	**more disc.**
PIC-01	PIC-06		.01	.002		.00	.846	
PIC-06	PIC-08		.00	.128		.00	.482	
PIC-07	PIC-07		**.04**	**<.001**	**easier**	.00	.026	
PIC-08	PIC-09		.00	.054		.00	.031	
PIC-09	PIC-18		**.03**	**<.001**	**harder**	.00	.076	
PIC-10	PIC-12	!	**.12**	**<.001**	**easier**	.00	.108	
PIC-11	PIC-14	!	**.26**	**<.001**	**harder**	**.01**	**<.001**	**less disc.**
PIC-14	PIC-24	!	**.03**	**<.001**	**harder**	**.01**	**.001**	**less disc.**
PIC-16	PIC-10		**.04**	**<.001**	**harder**	**.02**	**<.001**	**less disc.**
PIC-20	PIC-25		**.01**	**<.001**	**harder**	**.01**	**<.001**	**less disc.**

Note. These items were analyzed along with those in Table 3. ARR = Picture Arrangement, BD = Block Design, OBA = Object Assembly, PIC = Picture Completion. Item differences are marked ! for items not strictly identical but logically equivalent. R^2 is the Nagelkerke pseudo-R^2 from the logistic ordinal regression, *p* is the corresponding *p*-value based on Monte-Carlo simulations. Comparisons yielding significant DIF are in boldface.

Table 5

Comparison between WAIS-III and WAIS-IV for verbal subtests

Item ID		Item	Uniform DIF			Non-uniform DIF		
WAIS-III	WAIS-IV	differences	R^2	*p*-value	Direction	R^2	*p*-value	Direction
ARI-10	ARI-13		.00	.531		.00	.318	
COM-05	COM-06	s-	.00	.062		.00	.226	
COM-10	COM-04	s+	.00	.270		.00	.774	
COM-11	COM-13		**.03**	**<.001**	**harder**	.00	.185	
COM-13	COM-14		.00	.020		.00	.021	
DSF-04	DSF-04		**.01**	**<.001**	**harder**	.00	.112	
DSF-05	DSF-05		.00	.354		**.01**	**.001**	**more disc.**
DSF-06	DSF-06		.00	.736		.00	.084	
DSF-07	DSF-07		.00	.482		.00	.484	
DSF-08	DSF-08		.00	.329		.00	.424	
DSB-03	DSB-03		**.06**	**<.001**	**harder**	.00	.311	
DSB-04	DSB-04		**.05**	**<.001**	**harder**	.00	.104	
DSB-05	DSB-05		**.06**	**<.001**	**harder**	.00	.109	
DSB-06	DSB-06		**.07**	**<.001**	**harder**	.00	.744	
DSB-07	DSB-07		**.04**	**<.001**	**harder**	.00	.228	
INF-09	INF-09	s-	.00	.196		.00	.133	
INF-13	INF-12		**.01**	**<.001**	**harder**	.00	.777	
INF-18	INF-17		**.03**	**<.001**	**harder**	.00	.083	
INF-28	INF-25		**.04**	**<.001**	**harder**	.00	.233	
SIM-07	SIM-11	! s#	**.07**	**<.001**	**harder**	**.02**	**<.001**	**less disc.**
SIM-09	SIM-12	s-	.00	.003		.00	.045	
SIM-10	SIM-10	s-	.00	.006		.00	.021	
SIM-12	SIM-06	s-	**.07**	**<.001**	**easier**	**.01**	**<.001**	**more disc.**
SIM-19	SIM-16	s#	**.01**	**<.001**	**easier**	.00	.428	
VOC-08	VOC-09	s-	**.01**	**<.001**	**harder**	.00	.518	
VOC-15	VOC-21	s#	**.04**	**<.001**	**harder**	.00	.577	
VOC-18	VOC-13	s-	**.01**	**<.001**	**harder**	.00	.773	
VOC-20	VOC-18	s-	**.01**	**<.001**	**harder**	.00	.744	

Note. These items were analyzed along with those in Table 6. ARI =Arithmetic, COM = Comprehension, DSF = Digit Span Forward, DSB = Digit Span Backward, INF = Information, SIM = Similarities, VOC = Vocabulary. Item differences are marked ! for items not strictly identical but logically equivalent, or s+, s-, and s# for identical items with different scoring criteria in the more recent version (respectively more stringent criteria, more lenient criteria, and both more stringent and more lenient criteria). R^2 is the Nagelkerke pseudo-R^2 from the logistic ordinal regression, *p* is the corresponding *p*-value based on Monte-Carlo simulations. Comparisons yielding significant DIF are in boldface.

Table 6

Comparison between WAIS-III and WAIS-IV for visuo-spatial subtests

Item ID		Item	Uniform DIF			Non-uniform DIF		
WAIS-III	WAIS-IV	differences	R^2	*p*-value	Direction	R^2	*p*-value	Direction
BD-11	BD-11		**.01**	**<.001**	**easier**	.00	.460	
BD-12	BD-12		.00	.253		.00	.446	
BD-13	BD-13		.00	.102		.00	.452	
BD-14	BD-14		.00	.115		.00	.345	
MAT-08	MAT-08		**.04**	**<.001**	**easier**	.00	.288	
MAT-09	MAT-10		.00	.003		.00	.600	
MAT-10	MAT-11		.00	.090		.00	.016	
MAT-14	MAT-14		.00	.022		.00	.029	
MAT-15	MAT-15		.00	.055		.00	.371	
MAT-17	MAT-16		**.02**	**<.001**	**easier**	.00	.747	
MAT-22	MAT-19	!	**.02**	**<.001**	**easier**	.00	.036	
MAT-26	MAT-26		.00	.544		.00	.222	
PIC-07	PIC-04		.01	.003		**.01**	**<.001**	**less disc.**
PIC-08	PIC-07		.00	.076		.00	.034	
PIC-11	PIC-09		**.11**	**<.001**	**harder**	**.01**	**<.001**	**more disc.**
PIC-12	PIC-05		**.02**	**<.001**	**harder**	.00	.002	
PIC-16	PIC-08		**.01**	**<.001**	**easier**	**.01**	**.001**	**less disc.**
PIC-17	PIC-06		**.04**	**<.001**	**easier**	.00	.081	
PIC-19	PIC-19		**.50**	**<.001**	**harder**	.00	.510	
PIC-21	PIC-13		**.04**	**<.001**	**harder**	.00	.499	
PIC-22	PIC-10		**.03**	**<.001**	**easier**	.00	.009	
PIC-23	PIC-18		**.03**	**<.001**	**harder**	.00	.290	
PIC-24	PIC-16		.00	.377		.00	.416	
PIC-25	PIC-15		.00	.037		.00	.014	

Note. These items were analyzed along with those in Table 5. BD = Block Design, MAT = Matrix Reasoning, PIC = Picture Completion. Item differences are marked ! for items not strictly identical but logically equivalent. R^2 is the Nagelkerke pseudo-R^2 from the logistic ordinal regression, *p* is the corresponding *p*-value based on Monte-Carlo simulations. Comparisons yielding significant DIF are in boldface.