An examination of the openpsychometrics.org vocabulary test

Emil O. W. Kirkegaard, Ulster Institute for Social Research, United Kingdom, emil@emilkirkegaard.dk

Abstract

We examined data from the popular free online 45-item "Vocabulary IQ Test" from https://openpsychometrics.org/tests/VIQT/. We used data from native English speakers (n = 9,278). Item response theory analysis (IRT) showed that most items had substantial g-loadings (mean = .59, sd = .22), but that some were problematic (4 items being lower than .25). Nevertheless, we find that using the site's scoring rules (that include penalty for incorrect answers) give results that correlate very strongly (r = .92) with IRT-derived scores. This is also true when using nominal IRT. The empirical reliability was estimated to be about .90. Median test completion time was 9 minutes (median absolute deviation = 3.5) and was mostly unrelated to the score obtained (r = -.02).

The test scores correlated well with self-reported criterion variables educational attainment (r = .44) and age (r = .40). To examine the test for measurement bias, we employed both Jensen's method and differential item functioning (DIF) testing. With Jensen's method, we see strong associations with education (r = .89) and age (r = .88), and less so for sex (r = .32). With differential item functioning, we only tested the sex difference for bias. We find that some items display moderate biases in favor of one sex (13 items had $p_{bonferroni} < .05$ evidence of bias). However, the item pool contains roughly even numbers of male-favored and female-favored items, so the test level bias is negligible (|d| < 0.05). Overall, the test seems mostly well-constructed, and recommended for use with native English speakers.

Keywords: cognitive ability, intelligence, online testing, vocabulary, openpsychometrics.org, sex difference, measurement invariance, differential item functioning, Jensen's method, method of correlated vectors, sex bias

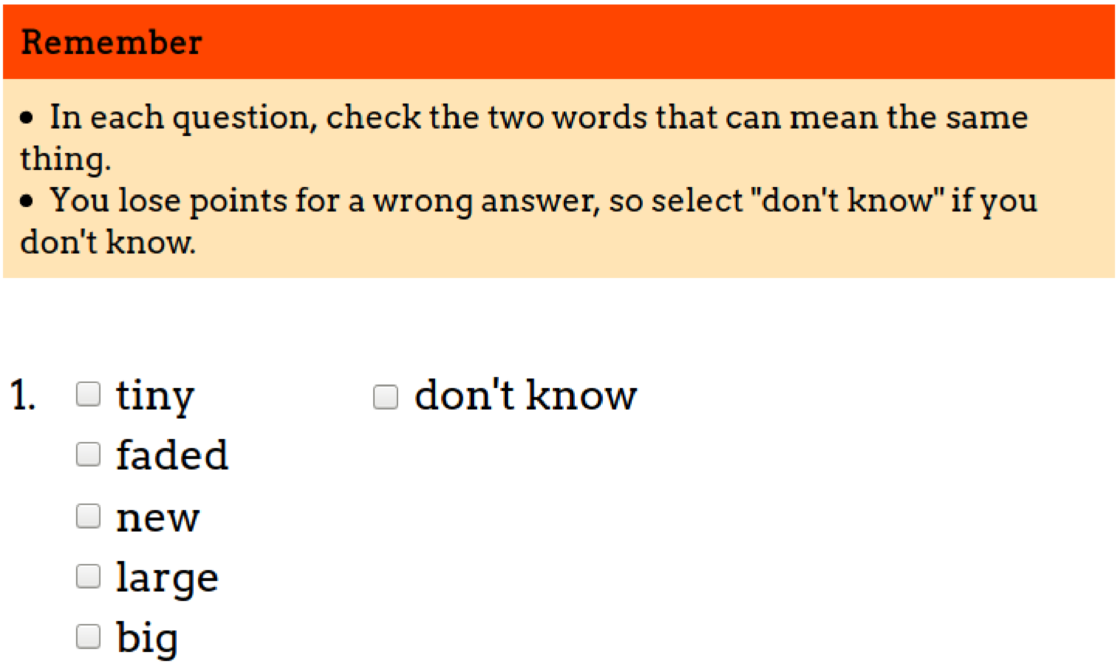
Introduction

Online psychological testing is popular. Unfortunately, there is a lack of validation of most online tests. This is also true for cognitive ability tests. The main exception is the ICAR (International Cognitive Ability Resource), which has seen extensive validation studies (Condon & Revelle, 2014; Dworak et al., 2020; Merz et al., 2020; Young et al., 2019).[1] Various national Mensa websites provide free figure reasoning tests (Raven-like) that provide IQ-normed results, but with unknown psychometric properties and norm data.[2] Here we examine a lesser known test simply called "Vocabulary IQ Test", which is available at https://openpsychometrics.org/tests/VIQT/. This is a 45-item multiple choice vocabulary test. The test origins or construction details are not given on the site, but this is presumably a newly developed test considering that the website brands itself as open

[1] The ICAR test is widely available for public use: https://discovermyprofile.com/tests/Intelligence/-/-. https://www.idrlabs.com/iq-16/test.php, https://www.sapa-project.org/.

[2] There is generally a Mensa group in each country, and many of them provide their own online screening or "for fun" tests. Examples: Denmark https://mensa.dk/iqtest/, Norway https://test.mensa.no/, Sweden https://www.mensa.se/bli-medlem/provtest-r1/, Romania https://mensaromania.ro/testari-mensa/test-online/.

source. The response format is the select-2-of-5 format, a somewhat unusual format (e.g., not covered in introduction books such as Kline, 2015). Figure X shows a screenshot of the test with the first item shown.



Remember

- In each question, check the two words that can mean the same thing.
- You lose points for a wrong answer, so select "don't know" if you don't know.

1. □ tiny □ don't know
□ faded
□ new
□ large
□ big

Figure X. First item and test instructions.

The purpose of the present study was to examine the psychometric properties of this test, as well as a limited exploration of the related data.

Data

Data for 12,173 persons are publicly available at the data page (https://openpsychometrics.org/_rawdata/). To reduce language bias, we only used data from persons who reported they were English native speakers. Inspection of the histogram of correct responses showed a small (<1%) lump of persons with near zero scores. These are users that click through the test for test purposes. We removed subjects with scores below 10 (less than 1%). The final sample had n = 9,278 subjects. Of these, 4,603 (49.6%) were female, and 4,286 (46.2%) were male. The remainer did not disclose their sex or reported “Other”. Aside from the 45 test items, the dataset also contains age, nationality, 25 items from a Big Five personality test, and the amount of time spent. Time spent was given in seconds. It had extreme skew (some people leave the browser tab open for days before completing it), and it was converted to minutes and winsorised to a maximum of 120 minutes. The personality data were not used in the present study.

All data and code output are available in the supplementary materials.

Results

The select-2-of-5 format of the data is mathematically equivalent to the select-1-of-10 format because there are 10 ways to pick 2 of the 5 options without duplication and order (i.e., (5*4)/2). By having people select two options, however, it is more space efficient than enumerating the pairwise options and having the subject read 10 response options. The site's approach to scoring the test is to convert the responses to dichotomous correct/incorrect format, and then sum the correct responses subtracted by the incorrect responses. A more advanced approach involves using item response theory (IRT) on the dichotomized items and then scoring the persons using the resulting model. However, a further refinement is to employ categorical/nominal IRT (Storme et al., 2019; Suh & Bolt, 2010). In this approach, each response is allowed to have its own relationship to the underlying trait. The benefit of this approach comes from the fact that the different distractors (incorrect response options) do not have the same expected trait levels, that is, some responses are more obviously incorrect than others, and this variation can be used for more precise or prediction scoring of persons given sufficient sample size (for a machine learning example, see Cutler et al., 2019). Here we scored the test data using 4 methods, 1) sum of correct responses, 2) sum of correct minus incorrect, 3) dichotomous/binary IRT using 2PL (2-parameter logistic), 4) categorical/nominal IRT using 2PLNRM (2-parameter logistic nominal response model) (Suh & Bolt, 2010).[3] The simple sum is the most commonly used method, and can be interpreted as a latent variable model with equal loadings (McNeish & Wolf, 2020). The advantage here is the simplicity of use, especially for manual scoring by hand, and the fact that one does not need to estimate factor loadings. Estimation of factor loadings in small samples produces unreliable results, and it may be better to simply assume equal loadings (Gorsuch, 2015; Ree et al., 1998). The simple sum with subtraction for incorrect responses attempts to deal with differences in guessing rates, by subtracting the expected score gains from this. This method should produce better estimates if all guessing is done completely at random and individuals simply vary in how much they guess. This assumption is not likely to be accurate, so it is unclear how this correction will affect estimates. The binary 2-parameter logistic model (2PL) allows for items to vary in difficulty and factor loading. Thus, items that are more informative for a subject are given more weight in the scoring, and there is no bias from the binary nature of the data. This model should produce more accurate estimates than the simple sum when items actually vary in factor loadings, which almost any collection of items will do to a large degree. The nominal nominal model further extends this by allowing that different incorrect responses may be differentially informative. In the binary models, each response is assumed to be informative only in two degrees, whether it is correct or incorrect. In the nominal, some incorrect responses are deemed more incorrect than others, and thus used to estimate the ability. This approach should be slightly more effective if a large sample is available for the model training (Storme et al., 2019).

In every case, the data were modeled as unidimensional. This score is best considered an approximation of the general intelligence factor (*g*) but with some influence by an orthogonal verbal ability. The IRT analyses were done using the **mirt** package for R (Chalmers et al., 2020) (MIRT = Multidimensional Item Response Theory). Table X shows the correlations between cognitive scores and criterion variables.

[3] We also tried other item models available in **mirt**'s *mirt()* function, namely nominal, graded, gpcm, and gpcmIRT. These all produced worse results than 2PLNRM. See the **mirt** documentation for details of implementation. https://rdrr.io/cran/mirt/man/mirt.html

	Sum score	**Sum score penalty**	**IRT binary**	**IRT cat**	**Education**	**Age**	**Time spent**
Sum score	1.00	0.95	0.97	0.96	0.43	0.38	-0.02
Sum score penalty	0.95	1.00	0.92	0.87	0.42	0.35	-0.02
IRT binary	0.97	0.92	1.00	0.99	0.44	0.40	-0.02
IRT cat	0.96	0.87	0.99	1.00	0.44	0.40	-0.02
Education	0.43	0.42	0.44	0.44	1.00	0.35	0.00
Age	0.38	0.35	0.40	0.40	0.35	1.00	0.02
Time spent	-0.02	-0.02	-0.02	-0.02	0.00	0.02	1.00

Table X. Correlations between four different variants of test scores and criterion variables IRT = item response theory.

The various scoring methods produced scores that were very strongly correlated, r’s .87 to .99. The two more advanced scoring methods produced slightly stronger correlations with the criterion variables. Notably, the penalty method produced the weakest results, perhaps due to being confounded with guessing strategies that are not much related to cognitive ability. Since the two IRT methods produced equivalent results, we chose the simpler binary version for further analysis.

In terms of reliability, Cronbach’s alpha was .90, and the empirical reliability of the IRT scores was also about .90 (.89 for binary IRT, .90 for categorical IRT; see *empirical_rxx()* function for details).

With regards to time spent, it is possible there could be nonlinear associations. Figure X shows the scatterplot.

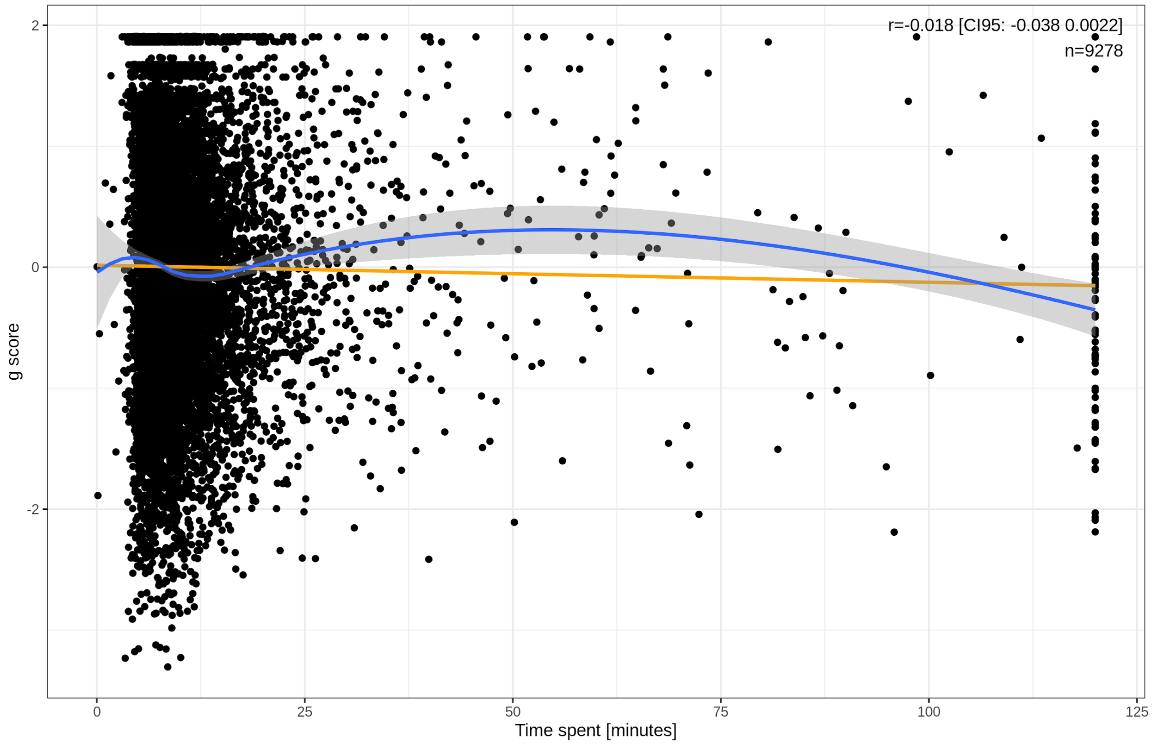


Figure X. Scatterplot of time spent and obtained score. Nonlinear fit provided by LOESS.[4]

While there was some evidence of a nonlinear non-monotonic trend, it was trivial in size. With regards to sex differneces, males obtained higher scores, as shown in Figure X.

[4] LOESS = locally estimated scatterplot smoothing, a method for deriving a moving average that will include nonlinear effects. This is the most common algorithm for handling the simple case of two continuous variables.

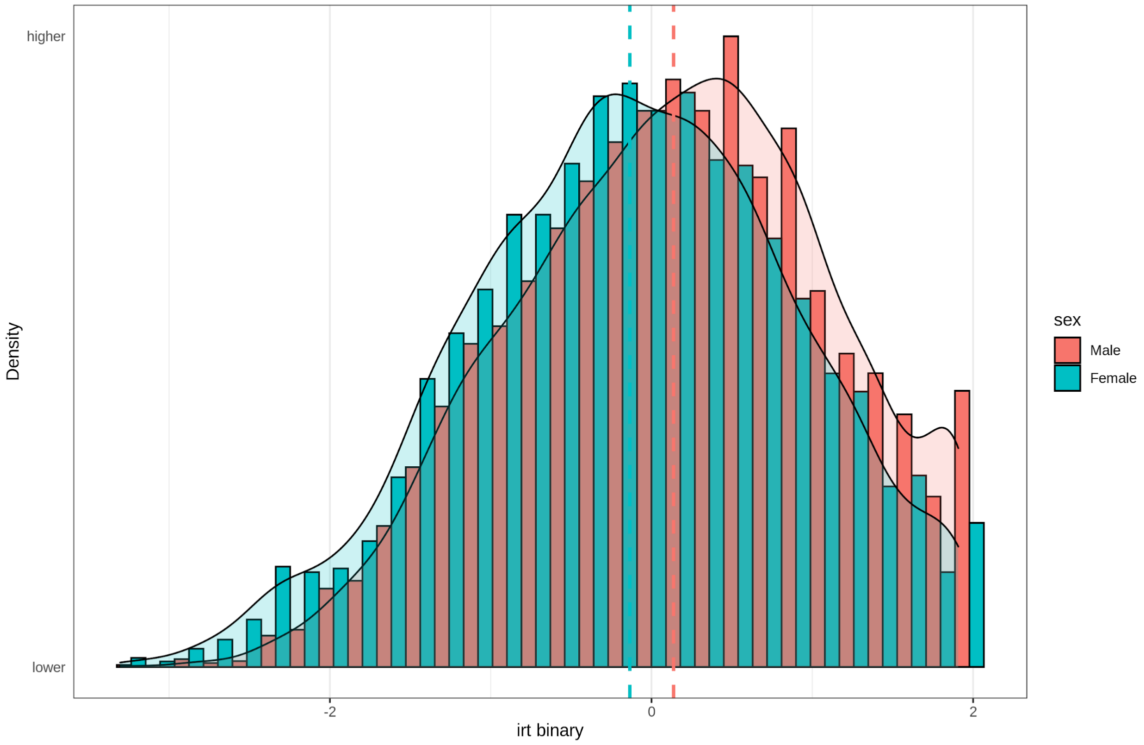


Figure X. Density-histogram of scores by sex.

Quantitatively speaking, the male advantage is 0.28 Cohen's d [95CI: -0.32 to -0.23, p < .0001]. While men had higher scores, women had higher dispersion, with standard deviations of 0.97 and 1.01, respectively. However, this female-advantage in dispersion may be a function of the test ceiling, as more men than women obtained perfect scores (3.1% vs. 1.6%, and 2.4% of all subjects). To examine whether some of this gap may be due to test bias, we carried out differential item functioning (DIF) testing using the functions provided by **mirt**.[5] This approach involves doing an initial leave-one-out run to look for items that show detectable DIF as compared to the other items as anchors. Then, in the second step, letting these items be freely estimated for each sex, using the remaining items as anchors (these are assumed to be unbiased). Finally, the total tests can be scored as scored using the invariant or partially invariant models (Meade, 2010) which will show the degree to which the items with bias impact the test scores. The results show negligible test level bias, with estimates of -0.04 and 0.03 (positive values indicate items that favor males), depending on a multiple testing adjustment (bonferroni) or not. Figure X shows the item functions.

[5] Specifically, we followed the approach by the package developer, as presented in two workshops (Chalmers, 2015a, 2015b). We emailed Chalmers in May 2020 to ask if the approach was still considered valid, and he affirmed that it is.

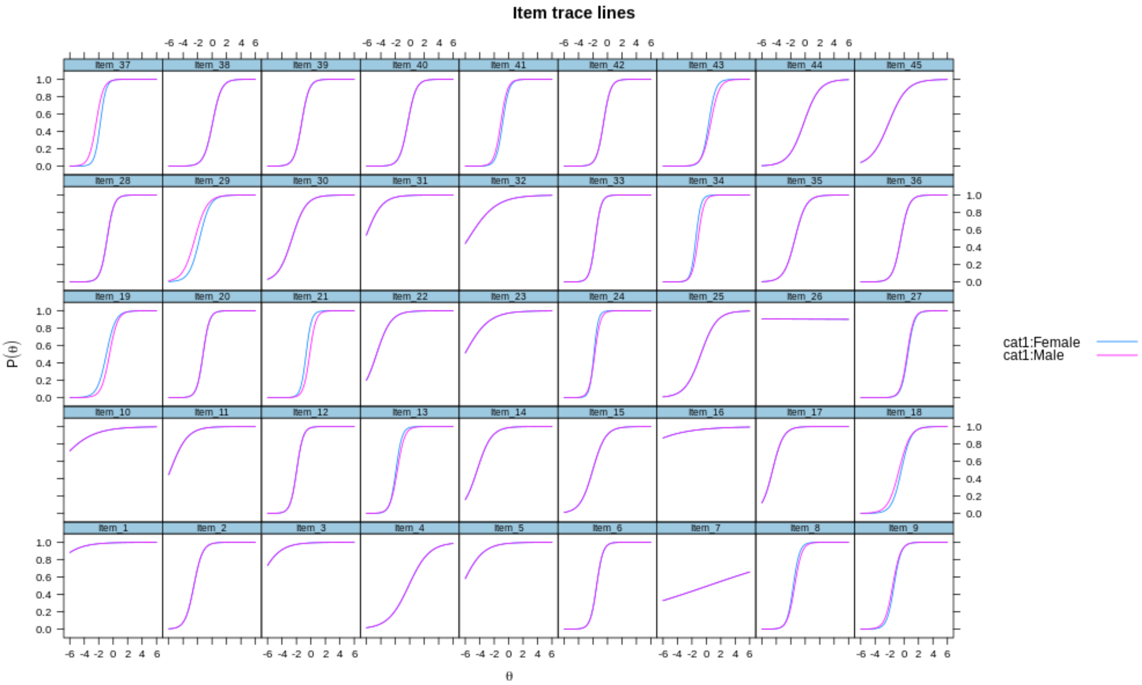


Figure X. Item response functions by sex.

It can be seen that some items are more informative than others (have a greater maximal slope), and that some show notable sex bias (when the lines are not overlapping, e.g., item 37 has male-bias and item 21 female-bias). Table X provides item-level information.

Item	Pass rate	Difficulty	Discrimination	Loading	Male d	Male bias	Age r	Education r	Time spent r
1	0.99	-5.13	0.51	0.29	-0.06	0.00	-0.06	0.05	-0.21
2	0.94	-3.87	1.66	0.70	0.36	0.00	0.24	0.20	-0.09
3	0.99	-4.99	0.67	0.37	-0.04	0.00	-0.01	0.06	-0.21
4	0.50	0.02	0.71	0.38	0.20	0.00	0.11	0.22	0.01
5	0.98	-4.29	0.70	0.38	0.04	0.00	0.10	0.06	-0.11
6	0.85	-2.96	2.25	0.80	0.24	0.00	0.31	0.38	-0.05
7	0.49	0.04	0.12	0.07	0.00	0.00	-0.09	-0.05	-0.02
8	0.85	-2.88	2.09	0.77	0.05	-0.23	0.34	0.38	-0.01
9	0.83	-2.47	1.95	0.75	0.33	0.12	0.33	0.40	-0.02
10	0.97	-3.40	0.40	0.23	0.21	0.00	-0.12	0.01	-0.05
11	0.99	-4.68	0.83	0.44	0.11	0.00	0.01	0.14	-0.07
12	0.92	-4.00	2.26	0.80	0.14	0.00	0.34	0.42	-0.02

13	0.90	-3.48	2.10	0.78	0.02	-0.25	0.20	0.22	-0.06
14	0.98	-4.18	1.00	0.51	0.15	0.00	0.08	0.09	-0.07
15	0.85	-2.09	1.13	0.55	0.35	0.00	0.25	0.25	-0.02
16	0.97	-3.53	0.29	0.17	0.07	0.00	-0.08	0.00	-0.03
17	0.99	-5.81	1.36	0.62	-0.07	0.00	0.06	0.16	0.01
18	0.58	-0.43	1.43	0.64	0.41	0.31	0.30	0.34	0.01
19	0.65	-0.90	1.59	0.68	-0.05	-0.42	0.40	0.36	0.02
20	0.80	-2.43	2.30	0.80	0.17	0.00	0.33	0.34	0.00
21	0.58	-0.60	2.26	0.80	-0.13	-0.50	0.40	0.39	-0.05
22	0.98	-4.39	1.01	0.51	0.17	0.00	0.08	0.24	0.01
23	0.97	-3.51	0.61	0.34	0.09	0.00	-0.04	0.04	0.02
24	0.92	-4.64	2.79	0.85	0.07	-0.18	0.33	0.38	-0.01
25	0.61	-0.52	0.91	0.47	0.27	0.00	0.13	0.28	-0.01
26	0.91	-2.27	-0.01	-0.01	0.05	0.00	-0.12	-0.07	-0.01
27	0.31	1.43	2.22	0.79	0.35	0.10	0.28	0.39	0.00
28	0.69	-1.35	2.09	0.78	0.32	0.00	0.42	0.45	-0.03
29	0.86	-2.31	1.33	0.61	0.47	0.45	0.12	0.23	0.00
30	0.89	-2.52	1.05	0.52	0.00	0.00	0.12	0.23	-0.04
31	0.99	-5.24	0.86	0.45	0.15	0.00	-0.03	0.01	0.00
32	0.94	-2.79	0.52	0.29	0.03	0.00	-0.06	0.05	-0.03
33	0.89	-3.71	2.49	0.83	0.29	0.00	0.40	0.46	0.00
34	0.82	-2.89	2.54	0.83	-0.03	-0.32	0.40	0.42	-0.03
35	0.77	-1.59	1.33	0.61	0.13	0.00	0.43	0.45	0.02
36	0.58	-0.47	1.80	0.73	0.12	0.00	0.41	0.41	0.00
37	0.92	-4.10	2.33	0.81	0.57	0.32	0.29	0.34	0.02
38	0.43	0.41	1.65	0.70	0.40	0.00	0.39	0.31	0.04
39	0.79	-2.14	1.96	0.76	0.12	0.00	0.33	0.36	0.00
40	0.52	-0.15	1.76	0.72	0.37	0.00	0.22	0.43	0.02
41	0.72	-1.66	2.21	0.79	0.46	0.23	0.24	0.32	-0.07
42	0.61	-0.76	2.08	0.77	0.22	0.00	0.37	0.37	-0.01

43	0.31	1.24	1.90	0.74	0.10	-0.20	0.34	0.37	0.00
44	0.50	0.01	0.94	0.48	0.19	0.00	0.20	0.31	0.03
45	0.80	-1.54	0.81	0.43	0.16	0.00	0.23	0.25	-0.02

Table X. Item statistics. _*r* means latent correlation with that variable (biserial; (Uebersax, 2015)). Male bias measured in Cohen's d.

Of the 45 items, not all are good items, as scored using the site's key. The mean g-loading is .59 (SD = 0.22). 4 items (7, 10, 16, and 26) had g-loadings below .25, and 1 below 0. These items should be revised or replaced.

Of the 45 items, 13 showed evidence of sex-bias ($p_{bonferroni}$ < .05). However, because the direction bias was symmetric around 0 (6 and 7 items), essentially no test level bias was seen. The distribution of item sex-bias is shown in Figure X.

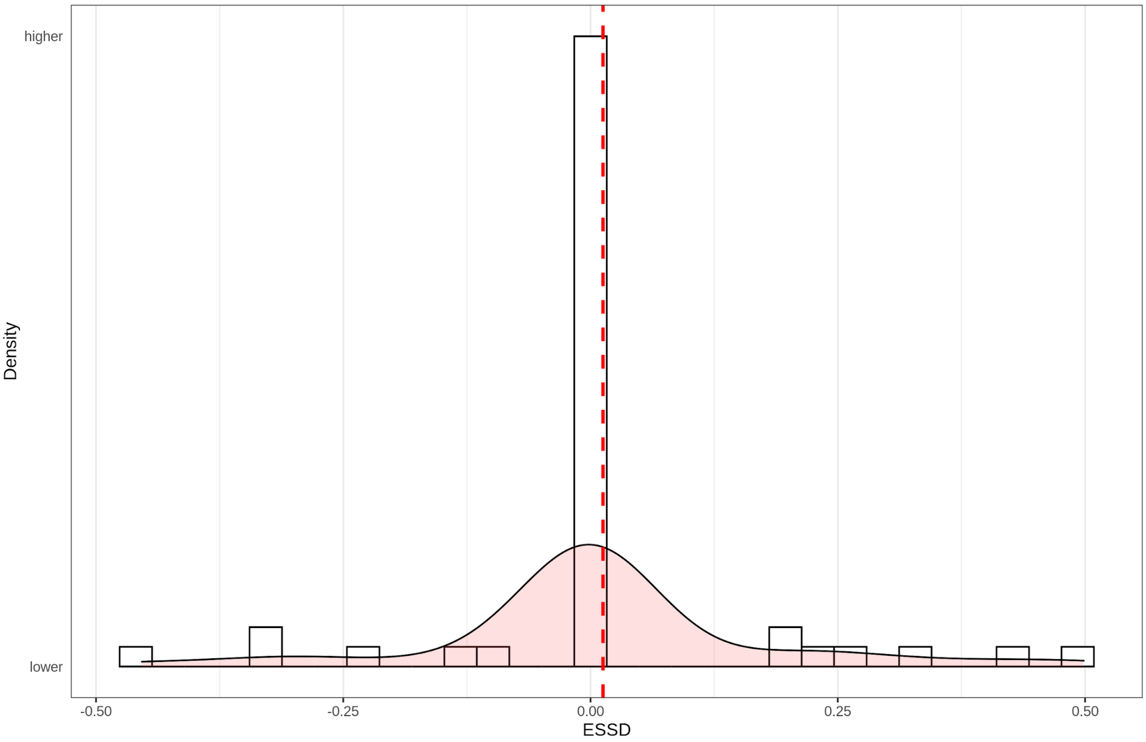


Figure X. Density-histogram of item sex-bias. The vertical line shows the mean.

Jensen's method (also called method of correlated vectors; (Dragt, 2010; Jensen, 1998)) is an alternative and simpler approach to examining the influence of latent variables. For any given scale, there are always a number of latent sources of variance, which may have different relationships to criterion variables. In the case of cognitive data, much research has been concerned with the relative influence of the general factor of intelligence (g) related to other sources of variance (non-g) (Fernandes et al., 2014; te Nijenhuis et al., 2014; te Nijenhuis & van der Flier, 2013; Michael A. Woodley of Menie et al., 2019). By theory, if *g* is the cause of the relationship between test scores and some criterion variable, then the items that are better measures of *g* should show stronger associations with that criterion variable. Figure X shows the scatterplots for the 4 criterion variables, and Table X shows the correlations between the item-level variables.

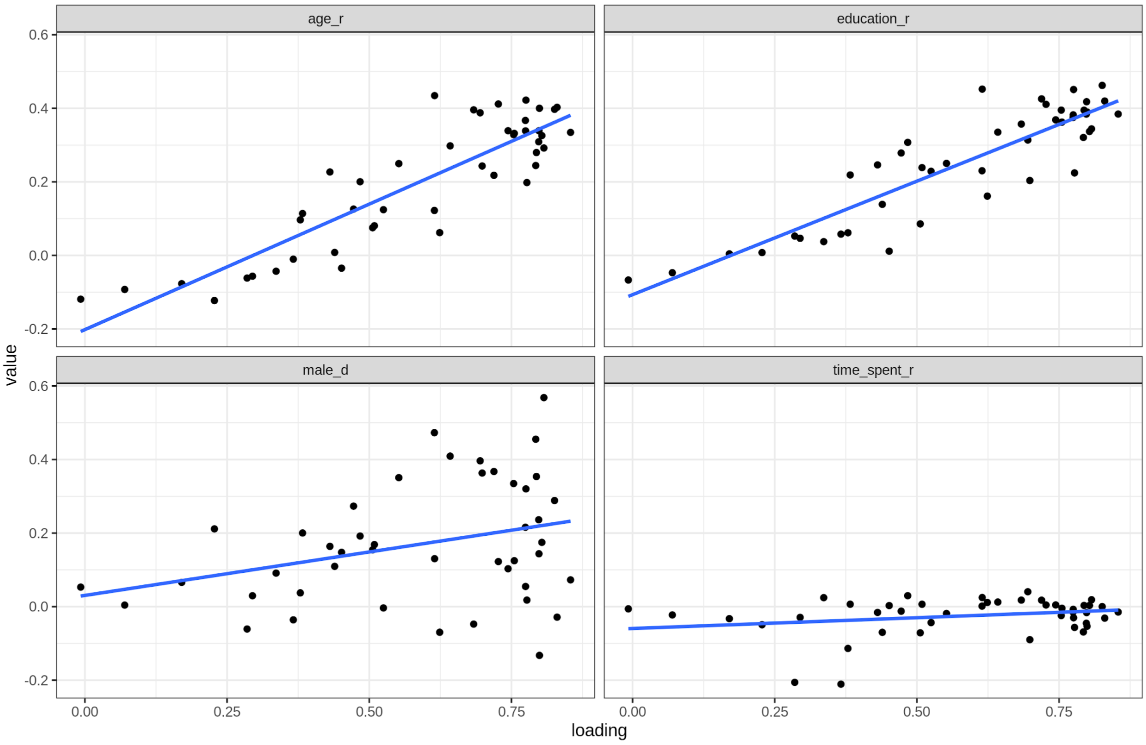


Figure X. Jensen’s method applied to 4 criterion variables. Correlations are .88, .89, .32, and .25, respectively, for age, education, male, and time spent.

	Pass rate	Difficulty	Discrimination	Loading	Male d	Male bias	Age r	Education r	Time spent r
Pass rate	1	-0.95 [-0.97 -0.91]	-0.23 [-0.49 0.07]	-0.25 [-0.51 0.05]	-0.24 [-0.50 0.06]	0.09 [-0.21 0.37]	-0.43 [-0.65 -0.16]	-0.45 [-0.66 -0.18]	-0.43 [-0.64 -0.15]
Difficulty	-0.95 [-0.97 -0.91]	1	0.13 [-0.17 0.40]	0.15 [-0.15 0.42]	0.26 [-0.03 0.52]	-0.05 [-0.34 0.25]	0.40 [0.12 0.62]	0.40 [0.12 0.62]	0.46 [0.19 0.67]
Discrimination	-0.23 [-0.49 0.07]	0.13 [-0.17 0.40]	1	0.97 [0.94 0.98]	0.27 [-0.02 0.52]	-0.17 [-0.44 0.13]	0.85 [0.75 0.92]	0.86 [0.76 0.92]	0.23 [-0.07 0.49]
Loading	-0.25 [-0.51 0.05]	0.15 [-0.15 0.42]	0.97 [0.94 0.98]	1	0.32 [0.03 0.56]	-0.13 [-0.41 0.17]	0.88 [0.79 0.93]	0.89 [0.81 0.94]	0.25 [-0.04 0.51]
Male d	-0.24 [-0.50 0.06]	0.26 [-0.03 0.52]	0.27 [-0.02 0.52]	0.32 [0.03 0.56]	1	0.71 [0.53 0.83]	0.27 [-0.02 0.52]	0.34 [0.05 0.57]	0.33 [0.04 0.57]

Male bias	0.09 [-0.21 0.37]	-0.05 [-0.34 0.25]	-0.17 [-0.44 0.13]	-0.13 [-0.41 0.17]	0.71 [0.53 0.83]	1	-0.22 [-0.48 0.08]	-0.13 [-0.41 0.17]	0.07 [-0.23 0.35]
Age r	-0.43 [-0.65 -0.16]	0.40 [0.12 0.62]	0.85 [0.75 0.92]	0.88 [0.79 0.93]	0.27 [-0.02 0.52]	-0.22 [-0.48 0.08]	1	0.94 [0.90 0.97]	0.35 [0.07 0.59]
Education r	-0.45 [-0.66 -0.18]	0.40 [0.12 0.62]	0.86 [0.76 0.92]	0.89 [0.81 0.94]	0.34 [0.05 0.57]	-0.13 [-0.41 0.17]	0.94 [0.90 0.97]	1	0.39 [0.11 0.61]
Time spent r	-0.43 [-0.64 -0.15]	0.46 [0.19 0.67]	0.23 [-0.07 0.49]	0.25 [-0.04 0.51]	0.33 [0.04 0.57]	0.07 [-0.23 0.35]	0.35 [0.07 0.59]	0.39 [0.11 0.61]	1

Table X. Item-level variables correlation matrix (45 items). Values in brackets are 95% confidence intervals.

The relationships for age and education are very strong. The relationship to male d is comparatively weaker, despite the results of the DIF analysis finding that the gap was not due to test bias. Our interpretation is that the biased items upset the relationship to the g-loadings. To test this, we carried out regression analysis and included the estimated bias from DIF. Results are shown in Table X.

Predictor/Model	**Simple**	**Add difficulty**	**Add bias**
Intercept	0.03 (0.066)	0.09 (0.076)	0.07 (0.043)
loading	0.24 (0.106)	0.21 (0.105)	0.28 (0.060***)
difficulty		0.02 (0.013)	0.02 (0.007**)
male_bias			0.77 (0.081***)
R2 adj.	0.083	0.111	0.717
N	45	45	45

Table X. Regression models for item analysis (Jensen’s method extended). Regression variables not standardized. * = p < .01, ** = p < .005, *** = p < .001.

The regression results confirm the hypothesis. The sex-biased items are outliers in the plot, and including their estimated bias results in a well fitting model (model adj. R2 = 72%). Figure X shows the item scatterplot with biased items marked.

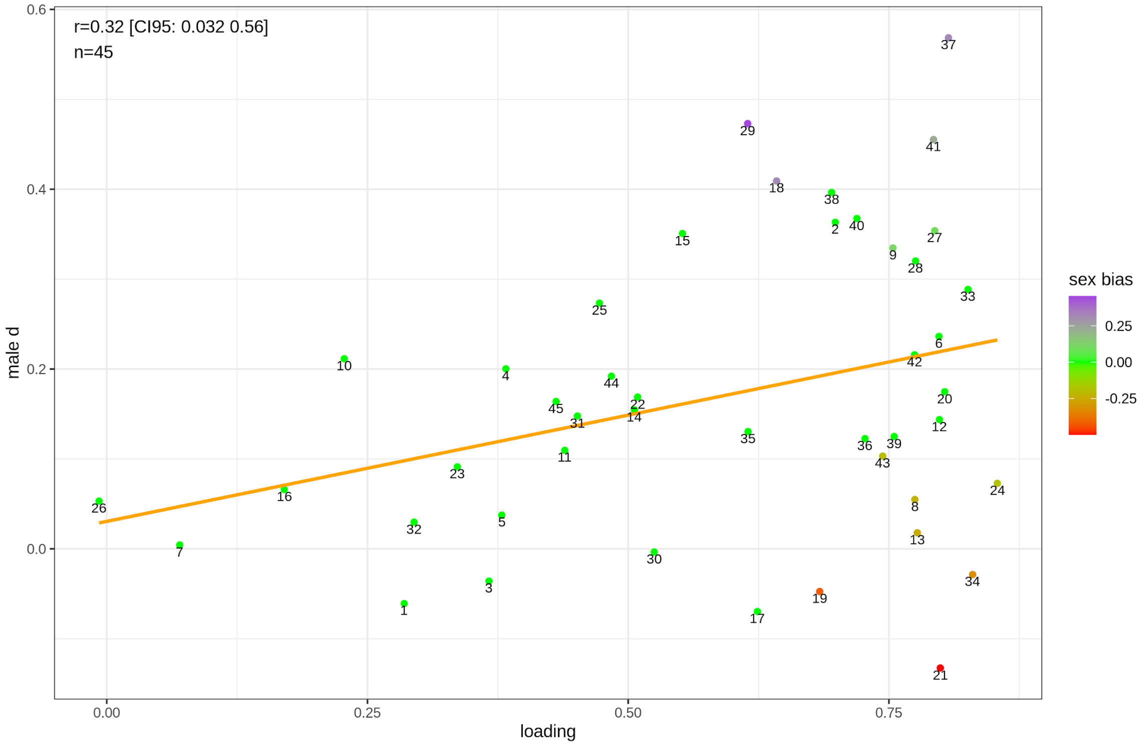


Figure X. Jensen's method on items for male advantage, with colors for DIF estimated item bias.

It can be seen in the plot that the outlying items with strong g-loadings are colored in the expected ways, with male-biased items above the regression line, and female-biased below.

Discussion

There were multiple findings of note. First, despite a few poor items, the test works well overall. The correlations to self-reported educational attainment and age were expected, as were the positive Jensen's method results for these (Dragt, 2010; Strenze, 2015). The reliability was good, estimated around .90 across methods. As such, the test can be recommended for public use. However, it should be noted that the norms are unknown. Under the assumption that they are based on the test takers, they are possibly inaccurate insofar as test takers are not representative of the general population. They may be smarter or duller, or have practiced the test more, or cheated by looking up the word definitions during the test (Cavanagh, 2014). To acquire better norm data, it is necessary to administer the test to a large representative population.

Second, we examined the test for sex bias using DIF testing. On this test, males obtained somewhat higher scores, d = 0.28 (4.2 IQ points). We found evidence of sex bias in 13 of the 45 items. However, the directions of bias were roughly balanced (6 and 7 items) such that the test level bias was near zero. This test can justifiably be used to compare scores of male and female examinees. We also employed the simpler Jensen's method approach and found that the results were congruent with the DIF testing results. Jensen's method showed a positive slope for g-loading and male advantage on an item, and when the effect of item bias was removed, the model fit very well (adj. R2 = 72%).

Jensen’s method yields very strong results for education and age, indicating older and more educated persons have greater vocabularies related to the general factor of the test. This finding is in line with prior results using test-level analysis (Dragt, 2010). With regards to Jensen’s method and item-level data, some prior studies have used suboptimal metrics (e.g. Al-Shahomee et al., 2017; Rushton J. Philippe et al., 2007), spawning a long list of critical papers (Wicherts, 2017, 2018a, 2018b; Wicherts & Johnson, 2009). Instead of using the difficulty, pass rates were used, which are nonlinear. Instead of g-loadings, item-whole point-biserial correlations were used, which are affected by the pass rate. Because of this, items with pass rates close to 0.50 have higher ‘g-loadings’, and these are the same items that have larger group gaps when measured in pass rates since a difference in latent ability of e.g. 1 d has the larger pass rate difference for an item when the overall pass rate is closest to 0.50. This confounding biases the resulting correlations in a positive direction. This present study did not use these faulty metrics and is thus unaffected by the criticism in those papers (see also Michael Anthony Woodley of Menie et al., 2020 for another study using this approach).

Per the above results, the male advantage we find cannot be explained by bias in the items. Though overall IQ scores usually favor males in adults (Lynn, 2017), vocabulary scores do not. Table X shows a comparison of large representative studies of adults on vocabulary tests.

Test	Country	Year	d	citation
meta-analysis of 40 studies	various	until 1988	0.02	(Hyde & Linn, 1988)
WAIS-3 vocabulary	USA	1997	0.04	(Chen & Lynn, 2020)
WAIS-3 vocabulary	Taiwan	2001	0.31	
WAIS-4 vocabulary	USA	2008	0.05	(Chen & Lynn, 2018)
WAIS-4 vocabulary	Taiwan	2015	0.20	
WAIS-4 vocabulary	Chile	2013	0.02	(Lynn, 2016)
WAIS-4 vocabulary	South Korea	2011	-0.01	(Lynn & Hur, 2016)
custom vocabulary test	Brazil	2014	-0.03	(Flores-Mendoza et al., 2016)

Table X. Summary of large representative studies of sex differences in vocabulary in adults.

The meta-analysis by Hyde & Linn included studies of children as well as adults, and the remaining studies included only adults and usually had more than 1000 subjects, mostly from standardization samples. Only the two studies from Taiwan show a male advantage of note, and the mean across all rows is 0.07 (without Taiwan, 0.02). (Lynn, 2021) carried out a meta-analysis of subtest results from Wechsler tests, and found similar results. For instance, there was a median male advantage of only 0.12 d on the vocabulary scale across 34 studies, less than half the male advantage we find in this study. While we don't know why we observe a notable difference where others generally don't, we speculate this may be due to a sex difference self-selection bias for online testing, such that duller men have a stronger tendency not to participate compared to duller women. It is known that academic study participation is related to intelligence and educational attainment, and that this effect differs by sex (Pirastu et al., 2021; but note that direction of bias varied between 23andme and UK Biobank!). However, it is not known how this generalizes to online tests taken at leisure.

Third, we find that the site's scoring approach of summing correct answers and subtracting the incorrect ones is inferior to using the simpler approach of summing correct answers only. Furthermore, using an IRT scoring approach is slightly superior to both of these simpler approaches (r with age .40 vs. .35/.38; r with education .44 vs. .42/.43). However, we find that using the full categorical data is not better than using the dichotomized data (Storme et al., 2019). It is thus suggested that the website also adopt a dichotomous IRT approach for scoring in conjunction with the sum of correct responses approach, given its ease of understanding. The current scoring rule that subtracts points for incorrect answers is suboptimal. The main limitation of this criterion analysis is that we only have 2 variables to investigate. It would be preferable to repeat this method comparison using a wider range of criterion variables, and preferably in a larger dataset, so that precision would be sufficiently high to detect even small differences between correlations (e.g., r = .20 vs. .22).

References

Al-Shahomee, A. A., Nijenhuis, J. te, Hoek, M. van den, Spanoudis, G., & Žebec, M. S. (2017). Spearman's Hypothesis Tested Comparing Young Libyan with European Children on the Items of the Standard Progressive Matrices. *Mankind Quarterly*, *57*(3). http://mankindquarterly.org/archive/issue/57-3/15

Cavanagh, T. (2014). Cheating on Online Assessment Tests: Prevalence and Impact on Validity. *Collected Faculty and Staff Scholarship*. https://scholar.dominican.edu/all-faculty/174

Chalmers, P. (2015a). *Multidimensional Item Response Theory Workshopin R*. https://philchalmers.github.io/mirt/extra/mirt-Workshop-2015_Day-1.pdf

Chalmers, P. (2015b). *Multidimensional Item Response Theory Workshopin R (day 2)*. https://philchalmers.github.io/mirt/extra/mirt-Workshop-2015_Day-2.pdf

Chalmers, P., Pritikin, J., Robitzsch, A., Zoltak, M., Kim, K., Falk, C. F., Meade, A., Schneider, L., King, D., Liu, C.-W., & Oguzhan, O. (2020). *mirt: Multidimensional Item Response Theory* (1.32.1) [Computer software]. https://CRAN.R-project.org/package=mirt

Chen, H.-Y., & Lynn, R. (2018). Sex Differences on the WAIS-IV in Taiwan and the United States. *Mankind Quarterly*, *59*(1). http://mankindquarterly.org/archive/issue/59-1/11

Chen, H.-Y., & Lynn, R. (2020). Sex Differences on the WAIS-III in Taiwan and the United States. *Mankind Quarterly*, *61*(2). http://mankindquarterly.org/upcoming/issue/61-2/9

Condon, D. M., & Revelle, W. (2014). The international cognitive ability resource: Development and initial validation of a public-domain measure. *Intelligence*, *43*, 52–64. https://doi.org/10.1016/j.intell.2014.01.004

Cutler, A., Dunkel, C. S., McLoughlin, S., & Kirkegaard, E. O. W. (2019). *Machine learning psychometrics: Improved cognitive ability validity from supervised training on item level data*. International Society for Intelligence Research, Minneapolis, MN, USA. https://www.researchgate.net/publication/334477851_Machine_learning_psychometrics_Improved_cognitive_ability_validity_from_supervised_training_on_item_level_data

Dragt, J. (2010). *Causes of group differences studied with the method of correlated vectors: A psychometric meta-analysis of Spearman's hypothesis*. http://dare.uva.nl/cgi/arno/show.cgi?fid=176083

Dworak, E. M., Revelle, W., Doebler, P., & Condon, D. M. (2020). Using the International Cognitive Ability Resource as an open source tool to explore individual differences in cognitive ability. *Personality and Individual Differences*, 109906. https://doi.org/10.1016/j.paid.2020.109906

Fernandes, H. B. F., Woodley, M. A., & te Nijenhuis, J. (2014). Differences in cognitive abilities among primates are concentrated on G: Phenotypic and phylogenetic comparisons with two meta-analytical databases. *Intelligence*, *46*, 311–322. https://doi.org/10.1016/j.intell.2014.07.007

Flores-Mendoza, C., Darley, M., & Fernandes, H. B. F. (2016). Cognitive Sex Differences in Brazil. *Mankind Quarterly*, *57*(1). https://doi.org/10.46469/mq.2016.57.1.4

Gorsuch, R. L. (2015). *Factor analysis* (Classic edition). Routledge, Taylor & Francis Group.

Hyde, J. S., & Linn, M. C. (1988). Gender differences in verbal ability: A meta-analysis. *Psychological Bulletin*, *104*(1), 53–69. https://doi.org/10.1037/0033-2909.104.1.53

Jensen, A. R. (1998). *The g factor: The science of mental ability*. Praeger.

Kline, P. (2015). *A Handbook of Test Construction (Psychology Revivals): Introduction to Psychometric Design*. Routledge. https://doi.org/10.4324/9781315695990

Lynn, R. (2016). Sex Differences on the WAIS-IV in Chile. *Mankind Quarterly*, *57*(1). https://doi.org/10.46469/mq.2016.57.1.5

Lynn, R. (2017). Sex Differences in Intelligence: The Developmental Theory. *Mankind Quarterly*, *58*(1). http://mankindquarterly.org/archive/issue/58-1/2

Lynn, R. (2021). Sex Differences in Verbal Abilities in the Wechsler Tests: A Review. *Mankind Quarterly*, *61*(3). http://mankindquarterly.org/archive/issue/61-3/16

Lynn, R., & Hur, Y.-M. (2016). Sex Differences on the WAIS-IV in the South Korean Standardization Sample. *Mankind Quarterly*, *57*(1). https://doi.org/10.46469/mq.2016.57.1.6

McNeish, D., & Wolf, M. G. (2020). Thinking twice about sum scores. *Behavior Research Methods*. https://doi.org/10.3758/s13428-020-01398-0

Meade, A. W. (2010). A taxonomy of effect size measures for the differential functioning of items and scales. *The Journal of Applied Psychology*, *95*(4), 728–743. https://doi.org/10.1037/a0018966

Merz, Z. C., Lace, J. W., & Eisenstein, A. M. (2020). Examining broad intellectual abilities obtained within an mTurk internet sample. *Current Psychology*.

https://doi.org/10.1007/s12144-020-00741-0

Pirastu, N., Cordioli, M., Nandakumar, P., Mignogna, G., Abdellaoui, A., Hollis, B., Kanai, M., Rajagopal, V. M., Parolo, P. D. B., Baya, N., Carey, C., Karjalainen, J., Als, T. D., Zee, M. D. van der, Day, F. R., Ong, K. K., Team, 23andMe Research, Consortium, iPSYCH, Morisaki, T., … Ganna, A. (2021). Genetic analyses identify widespread sex-differential participation bias. *BioRxiv*, 2020.03.22.001453. https://doi.org/10.1101/2020.03.22.001453

Ree, M. J., Carretta, T. R., & Earles, J. A. (1998). In Top-Down Decisions, Weighting Variables does Not Matter: A Consequence of Wilks' Theorem. *Organizational Research Methods*, *1*(4), 407–420. https://doi.org/10.1177/109442819814003

Rushton J. Philippe, Bons Trudy Ann, Vernon Philip A, & Čvorović Jelena. (2007). Genetic and environmental contributions to population group differences on the Raven's Progressive Matrices estimated from twins reared together and apart. *Proceedings of the Royal Society B: Biological Sciences*, *274*(1619), 1773–1777. https://doi.org/10.1098/rspb.2007.0461

Storme, M., Myszkowski, N., Baron, S., & Bernard, D. (2019). Same Test, Better Scores: Boosting the Reliability of Short Online Intelligence Recruitment Tests with Nested Logit Item Response Theory Models. *Journal of Intelligence*, *7*(3), 17. https://doi.org/10.3390/jintelligence7030017

Strenze, T. (2015). Intelligence and Success. In S. Goldstein, D. Princiotta, & J. A. Naglieri (Eds.), *Handbook of Intelligence* (pp. 405–413). Springer New York. http://link.springer.com/10.1007/978-1-4939-1562-0_25

Suh, Y., & Bolt, D. M. (2010). Nested Logit Models for Multiple-Choice Item Response Data. *Psychometrika*, *75*(3), 454–473. https://doi.org/10.1007/s11336-010-9163-7

te Nijenhuis, J., Jongeneel-Grimen, B., & Kirkegaard, E. O. W. (2014). Are Headstart gains on the g factor? A meta-analysis. *Intelligence*, *46*, 209–215. https://doi.org/10.1016/j.intell.2014.07.001

te Nijenhuis, J., & van der Flier, H. (2013). Is the Flynn effect on g?: A meta-analysis.

Intelligence, *41*(6), 802–807. https://doi.org/10.1016/j.intell.2013.03.001

Uebersax, J. S. (2015). *Introduction to the Tetrachoric and Polychoric Correlation Coefficients*. http://john-uebersax.com/stat/tetra.htm

Wicherts, J. M. (2017). Psychometric problems with the method of correlated vectors applied to item scores (including some nonsensical results). *Intelligence*, *60*, 26–38. https://doi.org/10.1016/j.intell.2016.11.002

Wicherts, J. M. (2018a). Ignoring psychometric problems in the study of group differences in cognitive test performance. *Journal of Biosocial Science*, *50*(6), 868–869. https://doi.org/10.1017/S0021932018000172

Wicherts, J. M. (2018b). This (method) is not fine. *Journal of Biosocial Science*, *50*(06), 872–874. https://doi.org/10.1017/S0021932018000184

Wicherts, J. M., & Johnson, W. (2009). Group differences in the heritability of items and test scores. *Proceedings of the Royal Society B: Biological Sciences*, *276*(1667), 2675–2683. https://doi.org/10.1098/rspb.2009.0238

Woodley of Menie, Michael A., te Nijenhuis, J., Shibaev, V., Li, M., & Smit, J. (2019). Are the effects of lead exposure linked to the g factor? A meta-analysis. *Personality and Individual Differences*, *137*, 184–191. https://doi.org/10.1016/j.paid.2018.09.005

Woodley of Menie, Michael Anthony, Kirkegaard, E. O. W., & Meisenberg, G. (2020). *Latent variable moderation of the negative fertility-item pass rate association in two, large datasets: An item response theory analysis.* https://doi.org/10.13140/RG.2.2.27203.96809

Young, S. R., Keith, T. Z., & Bond, M. A. (2019). Age and sex invariance of the International Cognitive Ability Resource (ICAR). *Intelligence*, *77*, 101399. https://doi.org/10.1016/j.intell.2019.101399