Submitted: 14th of March 2025 DOI: 10.26775/OP.2025.10.10
Published: 10th of October 2025 ISSN: 2597-324X

Black-White differences in an English vocabulary test using an online Prolific sample

Emil O. W. Kirkegaard [*] Meng Hu [†]

OpenPsych

Abstract

We sought to examine the Black-White difference in performance on a new English vocabulary test based on 225 items. Using data from the norm sample (N = 499, Prolific) we found a gap of $d = 0.74$. Adjusting for test reliability, this was $d = 0.75$ (reliability = .977). We examined measurement invariance using Differential Item Functioning (DIF). Biased items are flagged based on p-value < .05. We found 1 biased item after Bonferroni correction for multiple comparisons. An application of Jensen's Method of Correlated Vectors (MCV) to the item data showed a positive relationship between the Black-White difference and the items' factor loading, with a predicted gap of $d = 0.76$ at loading = 1. Findings were in line with prior research of minimal bias in vocabulary tests, and a *g*-related difference.

Key Words: Black-White gap, Cognitive Ability, Vocabulary, Test Bias, Differential Item Functioning, Spearman's Hypothesis, Method of Correlated Vectors

1 Introduction

There is a wealth of research and data on race differences in cognitive abilities in the USA, as well as elsewhere. Generally, research finds that African Americans (Blacks) score lower than European Americans (Whites), with a gap size that depends on the nature of the tests involved (Jensen, 1985). High quality, composite tests with representative samples usually find gaps of about 1 standard deviation, equivalent to 15 IQ points (Kirkegaard, 2022; Murray, 2021; Roth et al., 2001). This finding has been stable for many decades. Verbal subscales such as the 10-item vocabulary Wordsum test showed a smaller gap, about $d = 0.7$,[1] but no evidence of a gap narrowing over time (Hu, 2017, Table 7). One major study indicated a large reduction in the racial gaps (Dickens & Flynn, 2006) but the result was obtained from a projected trend line based on a small Black IQ gain per year. A closer look at the older birth cohorts from various data reveals that the secular narrowing of the Black-White IQ gap is rather ambiguous (Murray, 2006). More recent data analysis showed that the initial gap narrowed during the 1970s and 1980s but did not decline further after the late 1980s, remaining around 1 standard deviation (Murray, 2021). Another issue is that the secular score gains show a negative correlation with the *g* factor of intelligence (Dickens & Flynn, 2006, Table 2; te Nijenhuis & van der Flier, 2013, Table 2).

One line of research in group differences is whether the observed gaps can be explained by test bias, that is, whether measurement invariance is violated (Wicherts, 2016). A large amount of research on the topic shows that test bias rarely accounts for much of the observed gaps for non-immigrant, native language speaking minority groups (for a review, see: Hu, 2023) whereas bias has been reported for immigrants facing language barriers (Dolan et al., 2004; Wicherts & Dolan, 2010). Another line of research in group differences is whether the magnitude of the observed gaps across racial groups is a function of *g*-loadings, i.e., the test's or item's correlation with the *g* factor. This phenomenon has been called the Spearman's hypothesis (or sometimes Jensen Effect) and has been well supported using various methodologies, with the most common methods being

[*] Arthur Jensen Institute, Email: emil@emilkirkegaard.dk
[†] Independent researcher

[1] This estimate must be considered as lower-bound due to the low reliability of the Wordsum test (Cor et al., 2012).

Multi-Group Confirmatory Factor Analyses (Dolan, 2000; Dolan & Hamaker, 2001; Frisby & Beaujean, 2015; Hu et al., 2019; Kane & Oakland, 2010; Lasker et al., 2019, [2] 2021) and Jensen's Method of Correlated Vectors (Jensen, 1998; te Nijenhuis & Van den Hoek, 2016; te Nijenhuis et al., 2014, 2015a, 2015b, 2016a, 2017a, 2017b, 2019a, 2019b; Warne, 2016).

A point of contention with respect to testing Spearman's hypothesis (SH) comes from item analysis. Using the Method of Correlated Vectors (MCV), a series of studies (Rushton, 2002; Rushton & Skuy, 2000; Rushton et al. 2003) found a positive relationship between group differences in pass rates and item-total correlations (interpreted as *g*-loadings). But item-total correlations lack cross-sample consistency due to varying pass rates and suffer downward bias due to departure from an optimal 50/50 split in a dichotomous variable (Hunter & Schmidt, 2004, p. 36; Kirkegaard, 2015).

Furthermore, it has been demonstrated that substantial correlations can be easily obtained when the test at hand is completely different in one of the groups (Wicherts, 2017). This happens because item statistics based on Classical Test Theory (CTT) such as pass rates and item-total correlations are different between groups that differ in latent ability, whereas item parameters based on Item Response Theory (IRT) can be meaningfully compared across groups (Wicherts, 2017).

For this reason, we sought to replicate these previous findings on test bias and Spearman's hypothesis in a new online dataset of a recently developed vocabulary test by applying IRT methods.

2 Data & Methods

We used data from the American validation study of a new English vocabulary test (N = 499 for first-round data and N = 432 for follow-up data). While there were other groups in the dataset, the White and Black groups were the largest demographic subgroups which were large enough for analysis: 359 Whites and 63 Blacks. The base sample was recruited using the "representative sample" setting such that they reflect the US adult population in terms of age, sex, and race. The follow-up sample was recruited using prescreening to ensure only the participants in our base sample are eligible. For the first-round data collection, we created 159 vocabulary questions based on one single format: pick the correct synonym from a list of 5 words. However, our results based on the first wave showed that the test was too easy, with a mean pass rate of 77% and a non-trivial portion of participants with scores close to the ceiling. As a result, we constructed another set of 73 harder vocabulary items using two additional formats: choose 2 correct synonyms out of 5 words (composed of 37 items), choose 3 synonyms out of 5 words (composed of 32 items), and choose 1 correct synonym out of 5 words (composed of 4 items). Each item was scored as correct or incorrect depending on whether the subject chose the right option(s). The mean pass rate of this harder test was 41%. However, as in our previous study (Kirkegaard et al., 2024), we excluded items with poor loadings (<.25). Thus, for this study, the final test consists of 226 items[3] and the final sample consists of 422 participants (of which 63 were Blacks).

In our first wave, the pay rate was about 8 dollars per hour, which is the minimum according to Prolific rules. In the follow-up wave, we announced that the participants will be rewarded a 20% increase in their pay rate if they take the follow-up test, as we hoped to retain a maximum number of participants. The first wave of data was collected June 12th 2023 and the follow-up data was collected June 26th 2023. We did not close the follow-up test until two months elapsed as we accounted for the possibility that people would take holidays during summer. In both waves, we allowed participants to take the study using mobile, tablet, or desktop.

Our methodological strategy first involves calculating the IRT-estimated item loadings using the **mirt** package (Chalmers et al., 2019). From this, a general factor score is derived and the Black-White gap is calculated by using the IRT-based empirical internal reliability based on the function *empirical_rxx*() implemented in the **mirt** package. Then, we estimate the standardized gap in items' pass rate and employ Jensen's method based on our estimates of item loadings and item gaps. Finally, we conduct a Differential Item Functioning (DIF) analysis based on the IRT method implemented in the **mirt** package to evaluate whether item bias is present and consequential and whether there is a need to apply bias-correction procedures in our estimated gap.

[2] The result of their MGCFA analysis was displayed in full in their supplementary materials.

[3] We had to remove one item because it had no variance for Blacks, leaving us with 225 items.

3 Results

We first perform a factor analysis on the tetrachoric latent correlation matrix. This approach calculates item loadings on the same scale as test-level loadings, unaffected by item difficulty. Unlike item-test correlations, these loadings have the invariance property since they're not computed as within-group correlations based on sum scores. This addresses the issues raised by Wicherts (2017). The mean item loading was .67, with a minimum of .26. The first factor explained 47.5% of the variance. The IRT-based scores were computed from this model and age-normed to the White population (White mean/standard deviation of 100/15). The Black-White gap was 11.1 IQ, or *d* = 0.74 (Black standard deviation of 14.9). Given the extremely high reliability of the test (estimated r_{xx} = .977), the reliability adjusted gap was only marginally higher, *d* = 0.75. Figure 1 shows the distributions of vocabulary scores by racial group.

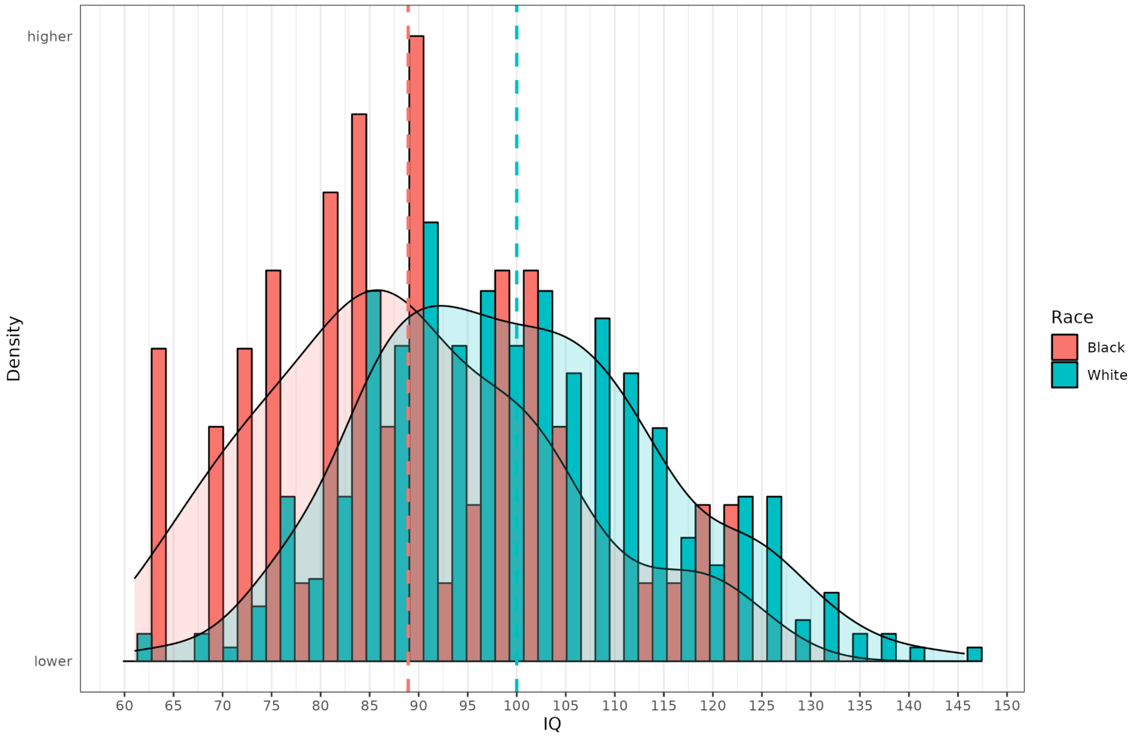

Figure 1. Distribution of vocabulary IQ scores for Black and White groups.

Jensen's MCV was applied to the item data. First, we derived the item standardized factor loadings. We computed the Cohen's *d* gap, estimated using the inverse normal distribution function (*qnorm()* in R) based on the pass rates by race. For instance, on the first item, Whites have a pass rate of 82.2% and Blacks of 68.3%. This corresponds to z score means of 0.922 and 0.475, yielding a gap of *d* = 0.447. Figure 2 shows the results.

A positive correlation was found (r = .54). If the present MCV result is to be compared with earlier reports, as displayed in Table 1, one would observe that the correlation between the Black-White gap and *g*-loading is consistently positive despite varying in magnitude.

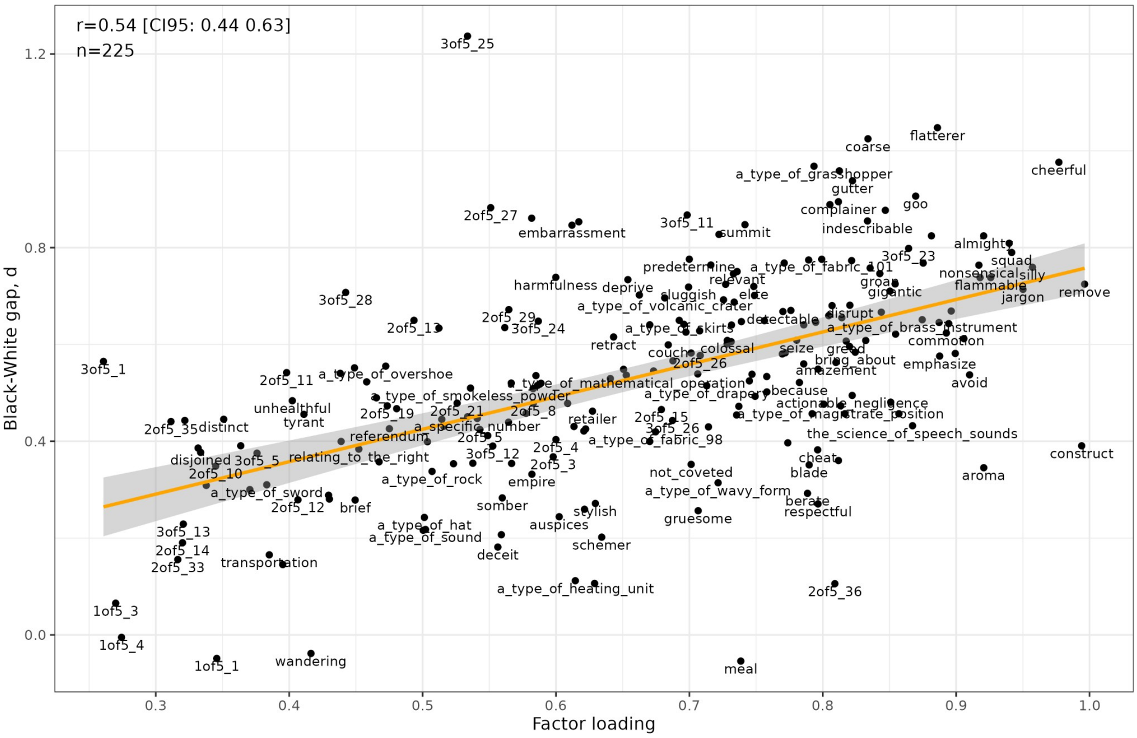

Figure 2. Jensen's MCV applied to 225 vocabulary items for the Black-White gap.

Table 1. Summary findings of Jensen's MCV for the Black-White gap at the item level[4]

	Study sample	Test type	BW gap at loading of 1.00[5]	r	Number of items	Sample size
Kirkegaard (2022). Study 1.	Prolific	Sentence Verification	.66	.22	24	915
Kirkegaard (2022). Study 2.	Prolific	Synonyms	.98	.56	10	772
Kirkegaard (2022). Study 3.	Prolific	Paper Folding	.73	.32	19	1239
Lasker et al. (2021)	VES	Verbal, Math & Fluid	1.35	.80	192	4179
Lasker et al. (2021)	NLSY79	Verbal & Math	1.23	.79	105	9126
This study	Prolific	Vocabulary	.76	.54	225	422

In this study, the Black sample was relatively small (N = 63) and the parameter estimates from IRT come with a certain degree of uncertainty. Thus, the true correlation will be larger as both sampling error and errors-in-variables problem bias the correlation towards 0 (Dutton & Kirkegaard, 2021). The estimated gap at a factor loading of 1.00 should be unbiased as random error in the outcome variable should not affect the slope, and the random error in the predictor variable is relatively small. Indeed, the estimated gap (intercept +

[4] These values are taken from Kirkegaard's analyses for these specific papers, accessible at: https://rpubs.com/EmilOWK/black_white_Prolific and https://rpubs.com/EmilOWK/VES_SH_items_2019.

[5] The predicted Black-White gap is obtained from a regression model, and is computed as the sum of the intercept and slope (here, the *g*-loading). In the case of Lasker et al. (2021), the model that includes difficulty and its interaction with *g* loading, as well as test type, is used.

slope$_{\text{g-loading}}$) at a perfect factor loading was 0.76 *d*, which closely matches both the observed gap (*d* = 0.74) and the reliability-adjusted value (*d* = 0.75).

Next we applied a DIF method to the two groups, which is performed using the **mirt** package (Chalmers et al., 2019). This program employs full-information maximum likelihood for estimating multiple group IRT models. The DIF procedure for purification, also called DIF screening, starts with a baseline model in which the latent means and latent variances are not constrained whereas the slopes and intercepts of all items are constrained to be equal across groups. From this baseline model, DIF statistics based on fit indices are computed by freeing (i.e., estimating) one item at a time while leaving other items constrained across groups. Non-invariance is typically determined based on goodness of fit and statistical significance.

In this analysis, we fit a partially invariant model based on items with p < .05 for the test, and scores based on this and the initial, baseline model are compared. The difference in the gap size is then computed. Since there are 225 items, the p values can be corrected for multiple testing using Bonferroni. Without Bonferroni testing, 27 items were found to be biased (27/225 = 12%, i.e., above chance expectation of 5%), but after multiple testing corrections only 1 item was biased.[6] In light of this latter result, one might reason that the conclusion of no bias is unwarranted because the finding of 27 significant items by chance alone is astronomically unlikely if all items were truly unbiased.[7] To address concerns that Bonferroni correction is overly conservative and relies on extreme-tail assumptions, we also applied the Benjamini-Hochberg (BH) procedure, which controls the False Discovery Rate (FDR). The BH correction yielded the same result, identifying only a single biased item. Crucially, regardless of the method, the impact on test-level bias was minimal (in Cohen's d, 0.02 against Blacks without correction, or 0.01 against Whites with Bonferroni correction). This means the statistical significance does not translate into practical significance.

4 Robustness test

Because the main analysis uses all items, including the 8 easy items that could have served as attention check to detect potentially poor responses. The supplemental analysis involves subsetting the sample to individuals who correctly answered all of the 8 easy items and removing the 8 easy items from all subsequent analyses. This leaves us with a total of 216 items and 441 participants (of which 327 are Whites and 47 are Blacks). Overall, the results from the main analyses are well replicated: the correlation between age and *g* did not change, the empirical reliabilities were virtually identical, and there was not a single item that displayed bias. There are however two discrepancies: the correlation between the Black-White gap in pass rate and item *g*-loading was down to .32 from .54 and the Black-White *d* gap was down to 0.57 from 0.74.

It can be argued that accounting for careless responding requires a causal justification, such as why careless responding is correlated with the predictor, or the outcome, or even group membership (Alsalti et al., 2025). The argument makes sense, but if low-effort responding differs across groups, then one group will display a greater nuisance factor (not intended to be measured). This not only threatens test validity but also creates spurious DIF. Although we removed participants who gave wrong answers to items with very obvious responses, this does not prove that they did not engage in careless responses in other items, as the wrong answer on a single (or even two) item(s) can be due to a mistap. Careless or low-effort response is better diagnosed through a series of tests aimed at detecting response patterns such as longstring responses, intra-individual variability, etc. (Arthur et al., 2021; DeSimone & Harms, 2018; Maniaci & Rogge, 2014). These tests could not be carried out because the answer options were randomized during the online test session (i.e., they appear in different order for each individual).

Another way to test for poor motivation is to model rapid-guessing behaviour through time response (Wise & Kong, 2005; Wise & Kingsbury, 2016; Wise & DeMars, 2010). This test could not be carried out either due to lacking data on time response for each individual item.

[6] Target word "meal", which of the following words is the synonym?: catatonia, elfin, bugaboo, hedonic, repast (correct). There is no obvious reason why this item should be biased, but it appeared to favor the Black group.

[7] Regarding the reviewer's concern on the extreme Bonferroni-corrected alpha ($\alpha = 0.05 / 225 = 0.00022$), the test statistic used was the Likelihood Ratio Test (LRT) statistic, which is asymptotically χ^2-distributed, not Gaussian. We agree that all asymptotic approximations, including the χ^2, must be treated with caution in the extreme tails.

5 Discussion

The result of the DIF analysis using IRT modeling, despite the small sample of Black participants, aligns with findings of no vocabulary test bias (Hu, 2023; Kirkegaard, 2022). The magnitude of the Black-White gap found in the present study (d = 0.75) is smaller than what was reported in another, recent online test (d = 0.99) on the same Prolific platform comprising a vocabulary subscale and a paper folding subscale (Kirkegaard, 2022). Crucially, both outcomes are likely attenuated by range restriction, as the Prolific platform self-selects against both low-IQ individuals (who may find the interface challenging) and high-IQ individuals (who are less attracted to low-paying tasks).

The result of the MCV analysis must be considered along with other tests of the Spearman's Hypothesis (SH). The most accepted method is the Multi-Group Confirmatory Analysis (MGCFA) based on which most studies have confirmed the hypothesis (Frisby & Beaujean, 2015; Hu et al., 2019; Kane & Oakland, 2010; Lasker et al., 2019; 2021), despite a few exceptions (Dolan, 2000; Dolan & Hamaker, 2001). Less common methods involve manipulating the g saturation of composite tests, McDaniel & Kepes (2014) found support for the hypothesis. SH is supported through the examination of Forward and Backward Digit Span, showing a BDS Black-White gap that is larger (d = 0.50) than the FDS gap (Jensen, 1998, p. 370). Perhaps the most powerful and direct way of testing SH is by examining the complexity of Elementary Cognitive Tasks (ECTs). Jensen (1998, p. 391) reported high correlation between task complexity and the magnitude of the Black-White gap (r = 0.86). Another finding supporting Spearman's Hypothesis comes from Dahlke & Sackett (2017) who performed a regression of group differences on cognitive saturation, defined as the correlation between cognitive and non-cognitive variables and cognitive ability.

Research on test bias is useful for examining the cultural hypothesis but also for vindicating the Spearman's Hypothesis. Not only for the purpose of valid cross-group comparison but also for estimating the true impact of g on group differences after accounting for item bias. There was indeed an account that the effect of g increases when biased items are removed (te Nijenhuis et al., 2016b).

A more convincing test of Spearman's hypothesis is to apply MCV in other subtests which are less subject to cultural influences than vocabulary. This would circumvent concerns about test g-loadings correlating with measures of cultural loadings (Malda et al., 2010; Kan et al., 2013; although see, te Nijenhuis & van der Flier, 2003; Hu, 2025). This concern is justified in the case of vocabulary, because verbal analogies based on highly familiar words tend to be more correlated with fluid abilities whereas analogies based on abstruse or specialized words tend to be more correlated with crystallized abilities (Jensen, 1980, p. 234). The vocabulary test used in this study contains a large number of items based on esoteric words, although our previous report showed that the higher the pass rate the higher the item factor loading (Kirkegaard et al., 2024). This should provide a counterweight to the criticism that g-loadings are ambiguous to interpret owing to their correlation with cultural loading, because in our vocabulary test the harder items were those using abstruse words, which are supposedly more culture loaded. Nevertheless, an ideal solution is to administer a complete battery of tests. In online settings however, extensive testing may result in fatigue and reduced motivation. This limitation can be overcome by the use of Computerized Adaptive Testing, which is based around the idea that a test can be shortened without any loss of measurement precision if the items actually match the examinee ability (Hambleton et al., 1991, ch. 9).

Given that Prolific participants are paid for taking the test but are not rewarded for performing well, a legitimate question is how payment affects motivation to do well in a test. Bates & Gignac (2022) conducted several analyses using the Prolific platform and found a modest effect size of (2.5 IQ) favoring the group who received extra payment for correct answers, a finding consistent with Gignac (2018) and Merritt et al. (2019). A potential threat to online surveys is differential validity stemming from group differences in motivation. If blacks have lower motivation than whites, resulting in careless responding, this could magnify the group differences. Indeed, careless or "lazy" responses should be treated cautiously (Arthur et al., 2021; DeSimone & Harms, 2018; Maniaci & Rogge, 2014) when causally justified (Alsalti et al., 2025). The **careless** package in R evaluates the response pattern using many indices to detect lazy responding such as longstring, even-odd consistency, psychometric synonyms/antonyms, intra-individual variability, Mahalanobis distance. Our data however is improper for conducting this kind of analysis since the order of answer options of every single item is randomized for participants.

Future research should explicitly measure motivation and its impact on test performance, particularly in online settings, to better understand its role in group differences. Such test-taking behavior has already been examined using response time data to account for rapid-guessing behavior (Michaelides et al., 2024; Lee & Jia, 2014), yet its impact on group differences is still understudied.

Supplementary Materials.

The supplementary materials can be accessed at: https://osf.io/t2j4s/
The results of the robustness tests can be accessed at: https://rpubs.com/MengHu/BWvocabProlific

References

Alsalti, T., Cummins, J., & Arslan, R. C. (2025). *Controlling for careless responding requires causal justification*. https://doi.org/10.31234/osf.io/zf79q_v2

Arthur, W., Hagen, E., & George, F. (2021). The Lazy or Dishonest Respondent: Detection and Prevention. *Annual Review of Organizational Psychology and Organizational Behavior*, *8*, 105–137. https://doi.org/10.1146/annurev-orgpsych-012420-055324

Bates, T. C., & Gignac, G. E. (2022). Effort impacts IQ test scores in a minor way: A multi-study investigation with healthy adult volunteers. *Intelligence*, *92*, 101652. https://doi.org/10.1016/j.intell.2022.101652

Chalmers, P., Pritikin, J., Robitzsch, A., Zoltak, M., Kim, K., Falk, C. F., Meade, A., Schneider, L., King, D., Liu, C.-W., & Oguzhan, O. (2019). *mirt: Multidimensional Item Response Theory (1.30) [Computer software]*. Accessed at: https://CRAN.R-project.org/package=mirt

Cor, M. K., Haertel, E., Krosnick, J. A., & Malhotra, N. (2012). Improving ability measurement in surveys by following the principles of IRT: The Wordsum vocabulary test in the General Social Survey. *Social Science Research*, *41*(5), 1003–1016. https://doi.org/10.1016/j.ssresearch.2012.05.007

Dahlke, J. A., & Sackett, P. R. (2017). The relationship between cognitive-ability saturation and subgroup mean differences across predictors of job performance. *Journal of Applied Psychology*, *102*(10), 1403–1420. https://doi.org/10.1037/apl0000234

DeSimone, J. A., & Harms, P. D. (2018). Dirty Data: The Effects of Screening Respondents Who Provide Low-Quality Data in Survey Research. *Journal of Business and Psychology*, *33*(5), 559–577. https://doi.org/10.1007/s10869-017-9514-9

Dickens, W. T., & Flynn, J. R. (2006). Black Americans reduce the racial IQ gap: Evidence from standardization samples. *Psychological Science*, *17*(10), 913–920.

Dolan, C. V. (2000). Investigating Spearman's hypothesis by means of multi-group confirmatory factor analysis. *Multivariate Behavioral Research*, *35*(1), 21–50. https://doi.org/10.1207/S15327906MBR3501_2

Dolan, C. V., & Hamaker, E. L. (2001). Investigating Black–White differences in psychometric IQ: Multi-group confirmatory factor analyses of the WISC-R and K-ABC and a critique of the method of correlated vectors. *Advances in Psychology Research*, *6*, 31–59.

Dolan, C. V., Roorda, W., & Wicherts, J. M. (2004). Two failures of Spearman's hypothesis: The GATB in Holland and the JAT in South Africa. *Intelligence*, *32*(2), 155–173. https://doi.org/10.1016/j.intell.2003.09.001

Dutton, E., & Kirkegaard, E. (2021). The Negative Religiousness–IQ Nexus is a Jensen Effect on Individual-Level Data: A Refutation of Dutton et al.'s 'The Myth of the Stupid Believer.' *Journal of Religion and Health*. https://doi.org/10.1007/s10943-021-01351-1

Frisby, C. L., & Beaujean, A. A. (2015). Testing Spearman's hypotheses using a bi-factor model with WAIS-IV/WMS-IV standardization data. *Intelligence*, *51*, 79–97. https://doi.org/10.1016/j.intell.2015.04.007

Gignac, G. E. (2018). A moderate financial incentive can increase effort, but not intelligence test performance in adult volunteers. *British Journal of Psychology*, *109*(3), 500–516. https://doi.org/10.1111/bjop.12288

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage Press.

Hu, M. (2017). An update on the secular narrowing of the Black–White gap in the Wordsum vocabulary test (1974–2012). *Mankind Quarterly*, *58*(2), 324–354. https://doi.org/10.46469/mq.2017.58.2.11

Hu, M. (2023). On The Validity of The GSS Vocabulary Test Across Groups. *OpenPsych*. https://doi.org/10.26775/OP.2023.06.22

Hu, M. (2025). Spearman's g Explains Black–White but not Sex Differences in Cognitive Abilities in the Project Talent *OpenPsych*. https://doi.org/10.26775/OP.2025.07.18

Hu, M., Lasker, J., Kirkegaard, E. O. W., & Fuerst, J. G. (2019). Filling in the gaps: The association between intelligence and both color and parent-reported ancestry in the National Longitudinal Survey of Youth 1997. *Psych*, *1*(1), 240–261. https://doi.org/10.3390/psych1010017

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*. Thousand Oaks, CA: SAGE Publications, Inc.

Jensen, A. R. (1980). *Bias in Mental Testing*. New York: Free Press.

Jensen, A. R. (1985). The nature of the Black–White difference on various psychometric tests: Spearman's hypothesis. *Behavioral and Brain Sciences*, *8*(2), 193–219. https://doi.org/10.1017/S0140525X00020392

Jensen, A. R. (1998). *The g Factor: The Science of Mental Ability*. Westport, CT: Praeger.

Kan, K. J., Wicherts, J. M., Dolan, C. V., & van der Maas, H. L. (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

Kane, H. D., & Oakland, T. D. (2010). Group Differences in Cognitive Ability: A CHC Theory Framework. *Mankind Quarterly*, *50*(4), 318–331. https://doi.org/10.46469/mq.2010.50.4.4

Kirkegaard, E. O. (2015). Spearman's hypothesis on item-level data from Raven's Standard Progressive Matrices: A replication and extension. *The Winnower*.

Kirkegaard, E. O. (2022). The Intelligence Gap between Black and White Survey Workers on the Prolific Platform. *Mankind Quarterly*, *63*(1), 79–88. https://doi.org/10.46469/mq.2022.63.1.3

Lasker, J., Nyborg, H., & Kirkegaard, E. O. W. (2021). Spearman's Hypothesis in the Vietnam Experience Study and National Longitudinal Survey of Youth '79. https://doi.org/10.31234/osf.io/m4yn9

Lasker, J., Pesta, B. J., Fuerst, J. G., & Kirkegaard, E. O. (2019). Global ancestry and cognitive ability. *Psych*, *1*(1), 431–459. https://doi.org/10.3390/psych1010034

Lee, Y. H., & Jia, Y. (2014). Using response time to investigate students' test-taking behaviors in a NAEP computer-based study. *Large-scale Assessments in Education*, *2*(8), 1–24. https://doi.org/10.1186/s40536-014-0008-1

Malda, M., van de Vijver, F. J. R., & Temane, Q. M. (2010). Rugby versus Soccer in South Africa: Content familiarity contributes to cross-cultural differences in cognitive test scores. *Intelligence*, *38*(6), 582–595. https://doi.org/10.1016/j.intell.2010.07.004

Maniaci, M. R., & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality*, *48*, 61–83. https://doi.org/10.1016/j.jrp.2013.09.008

McDaniel, M. A., & Kepes, S. (2014). An Evaluation of Spearman's Hypothesis by Manipulating g Saturation. *International Journal of Selection and Assessment*, *22*(4), 333–342. https://doi.org/10.1111/ijsa.12081

Merritt, V. C., Rabinowitz, A. R., Guty, E., Meyer, J. E., Greenberg, L. S., & Arnett, P. A. (2019). Financial incentives influence ImPACT validity indices but not cognitive composite scores. *Journal of Clinical and Experimental Neuropsychology*, *41*(3), 312–319. https://doi.org/10.1080/13803395.2018.1551519

Michaelides, M. P., Ivanova, M. G., & Avraam, D. (2024). The impact of filtering out rapid-guessing examinees on PISA 2015 country rankings. *Psychological Test and Assessment Modeling*, *66*, 50–62. https://doi.org/10.2440/001-0012

Murray, C. (2006). Changes over time in the Black–White difference on mental tests: Evidence from the children of the 1979 cohort of the National Longitudinal Survey of Youth. *Intelligence*, *34*(6), 527–540. https://doi.org/10.1016/j.intell.2006.07.004

Murray, C. (2021). *Facing Reality: Two Truths about Race in America*. New York: Encounter Books.

Roth, P. L., Bevier, C. A., Bobko, P., Switzer, F. S., & Tyler, P. (2001). Ethnic group differences in cognitive ability in employment and educational settings: A meta-analysis. *Personnel Psychology*, *54*(2), 297–330. https://doi.org/10.1111/j.1744-6570.2001.tb00094.x

Rushton, J. P. (2002). Jensen effects and African/Coloured/Indian/White differences on Raven's Standard Progressive Matrices in South Africa. *Personality and Individual Differences*, *33*(8), 1279–1284. https://doi.org/10.1016/S0191-8869(02)00012-0

Rushton, J. P., & Skuy, M. (2000). Performance on Raven's Matrices by African and White university students in South Africa. *Intelligence*, *28*(4), 251–265. https://doi.org/10.1016/S0160-2896(00)00035-0

Rushton, J. P., Skuy, M., & Fridjhon, P. (2003). Performance on Raven's Advanced Progressive Matrices by African, East Indian, and White engineering students in South Africa. *Intelligence*, *31*(2), 123–137. https://doi.org/10.1016/S0160-2896(02)00140-X

te Nijenhuis, J., Al-Shahomee, A. A., van den Hoek, M., Allik, J., Grigoriev, A., & Dragt, J. (2015a). Spearman's hypothesis tested comparing Libyan secondary school children with various other groups of secondary school children on the items of the Standard Progressive Matrices. *Intelligence*, *50*, 118–124. https://doi.org/10.1016/j.intell.2015.03.002

te Nijenhuis, J., Bakhiet, S. F., van den Hoek, M., Repko, J., Allik, J., Žebec, M. S., Sukhanovskiy, V., & Abduljabbar, A. S. (2016a). Spearman's hypothesis tested comparing Sudanese children and adolescents with various other groups of children and adolescents on the items of the Standard Progressive Matrices. *Intelligence*, *56*, 46–57. https://doi.org/10.1016/j.intell.2016.02.010

te Nijenhuis, J., Batterjee, A. A., Van Den Hoek, M., Allik, J., & Sukhanovskiy, V. (2017a). Spearman's hypothesis tested comparing Saudi Arabian children and adolescents with various other groups of children and adolescents on the items of the Standard Progressive Matrices. *Journal of Biosocial Science*, *49*(5), 634–647. https://doi.org/10.1017/S0021932016000535

te Nijenhuis, J., Choi, Y. Y., van den Hoek, M., Valueva, E., & Lee, K. H. (2019a). Spearman's hypothesis tested comparing Korean young adults with various other groups of young adults on the items of the Advanced Progressive Matrices. *Journal of Biosocial Science*, *51*(6), 875–912. https://doi.org/10.1017/S0021932019000026

te Nijenhuis, J., David, H., Metzen, D., & Armstrong, E. L. (2014). Spearman's hypothesis tested on European Jews vs non-Jewish Whites and vs Oriental Jews: Two meta-analyses. *Intelligence*, *44*, 15–18. https://doi.org/10.1016/j.intell.2014.02.002

te Nijenhuis, J., & van den Hoek, M. (2016). Spearman's hypothesis tested on Black adults: A meta-analysis. *Journal of Intelligence*, *4*(2), 6. https://doi.org/10.3390/jintelligence4020006 te Nijenhuis, J., van den Hoek, M., & Armstrong, E. L. (2015b). Spearman's hypothesis and Amerindians: A meta-analysis. *Intelligence*, *50*, 87–92. https://doi.org/10.1016/j.intell.2015.02.006

te Nijenhuis, J., van den Hoek, M., & Dragt, J. (2019b). A meta-analysis of Spearman's hypothesis tested on Latin-American Hispanics, including a new way to correct for imperfectly measuring the construct of g. *Psych*, *1*(1), 101–122. https://doi.org/10.3390/psych1010008

te Nijenhuis, J., van den Hoek, M., & Willigers, D. (2017b). Testing Spearman's hypothesis with alternative intelligence tests: A meta-analysis. *Mankind Quarterly*, *57*(4), 687–705. https://doi.org/10.46469/mq.2017.57.4.12

te Nijenhuis, J., & van der Flier, H. (2003). Immigrant–majority group differences in cognitive performance: Jensen effects, cultural effects, or both? *Intelligence*, *31*(5), 443–459. https://doi.org/10.1016/S0160-2896(03)00027-8

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

te Nijenhuis, J., Willigers, D., Dragt, J., & van der Flier, H. (2016b). The effects of language bias and cultural bias estimated using the method of correlated vectors on a large database of IQ comparisons between native Dutch and ethnic minority immigrants from non-Western countries. *Intelligence*, *54*, 117–135. https://doi.org/10.1016/j.intell.2015.12.003

Warne, R. T. (2016). Testing Spearman's hypothesis with Advanced Placement Examination data. *Intelligence*, *57*, 87–95. https://doi.org/10.1016/j.intell.2016.05.002

Wicherts, J. M. (2016). The importance of measurement invariance in neurocognitive ability testing. *The Clinical Neuropsychologist*, *30*(7), 1006–1016. https://doi.org/10.1080/13854046.2016.1205136

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., & Dolan, C. V. (2010). Measurement invariance in confirmatory factor analysis: An illustration using IQ test performance of minorities. *Educational Measurement: Issues and Practice*, *29*(3), 39–47. https://doi.org/10.1111/j.1745-3992.2010.00182.x

Wise, S. L., & DeMars, C. E. (2010). Examinee noneffort and the validity of program assessment results. *Educational Assessment*, *15*(1), 27–41. https://doi.org/10.1080/10627191003673216

Wise, S. L., & Kingsbury, G. G. (2016). Modeling student test-taking motivation in the context of an adaptive achievement test. *Journal of Educational Measurement*, *53*(1), 86–105. https://doi.org/10.1111/jedm.12102

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*(2), 163–183. https://doi.org/10.1207/s15324818ame1802_2