On The Validity of The GSS Vocabulary Test Across Groups

Meng Hu*

*Correspondence: mh19870410@gmail.com

Abstract

The psychometric properties of the Wordsum vocabulary test across race and gender groups has not been studied yet. Taking advantage of a large sample of American adults from the General Social Survey (GSS), the Differential item/test functioning (DIF/DTF) were evaluated across gender and racial groups by using Item Response Theory (IRT). Two items displayed DIF with respect to race (as blacks/whites) whereas four items displayed DIF with respect to gender. However, because the DIFs run in both directions, there is no consistent bias against either group at the test level. Despite being culturally loaded, the Wordsum shows no evidence of culture bias.

Key Words: Intelligence, Vocabulary, Wordsum, Group Differences, Differential Item Functioning, Test Bias

1. Introduction

Critics of early tests and intelligence testing pointed out the possible flaws of the tests when used with students from racial/ethnic minority groups. This debate was particularly salient in the United States and prompted many discussions about test bias (Jensen, 1980). Culture bias, the most commonly held argument, historically assumed that ethnic minorities such as blacks and hispanics score lower than whites because cognitive tests measure knowledge specific to the white culture. As Jensen (1980, ch. 14) mentioned however, there is no way one could draw a strong line between a culture-loaded and culture-free test. Instead, one should think of cultural distance. For instance, American people regardless of race and age will all correctly answer a question such as “What are the colors of the American flag” but people living in a secluded, isolated country or area in the world will all likely get it wrong.

There is a widespread belief that psychometric bias found in a test or test item necessarily runs against low-performing minority groups. As statistical methods showed however, biases often run in both directions and sometimes even cancel out. Indeed, psychometricians designed statistical tests for detecting questions (i.e., items) that are potentially biased with respect to group differences, known as Differential Item Functioning (DIF) methods. The DIF can take on multiple forms. If the group difference in item response is homogeneous across all levels of ability, the item shows uniform DIF. But when the group difference varies across the levels of ability, the item shows non-uniform DIF. A more complex version of non-uniform DIF is the crossing DIF, in which one group is advantaged at some levels of ability but disadvantaged at some other levels of ability. DIF methods can be classified into two categories: observed variable and latent variable methods. In both types of methods, test fairness is achieved when the two groups do not differ in item response probabilities after being equated at the total score level (either observed or latent score). The manifest variable approach includes the Partial Correlation (PC), Mantel-Haenszel (MH), Logistic Regression

(LR), Standardization (STD), whereas the latent variable approach includes the Simultaneous Item Bias (SIBTEST), Item Response Theory (IRT), Multi-Group Confirmatory Factor Analysis (MGCFA). The great majority of test bias comparing blacks and whites reach an agreement about the fairness of cognitive and achievement tests.

Studies using the Partial Correlation were numerous but old. Stricker (1982) found DIFs in the GRE verbal test which were evenly distributed across racial groups and gender groups. Reynolds et al. (1984) found in the PPVT few more items biased against blacks but of very small magnitude. Willson et al. (1989) found in the K-ABC more DIFs against blacks than whites but of little practical consequence. Conoley (2003) analyzed the black-white and hispanic-white differences in the PPVT-III and reported no consistent bias against blacks but a bias favoring hispanics.

Many studies also employed the Mantel-Haenszel. Scheuneman & Gerritz (1990) analyzed the SAT verbal and GRE verbal but while the SAT was minimally biased against whites the GRE was minimally biased against blacks. Nandakumar et al. (1993) examined the 60 items of the GATSB, a test measuring differential behaviors that would affect results obtained on the WISC-III standardization, in which only 5 items were biased against whites, 1 against hispanics and 1 against blacks in the racial group analysis and 1 biased against boys in the gender group analysis. Roznowski & Reith (1999) found no overall bias against blacks on a variety of aptitude tests such as vocabulary, basic arithmetic, geometry & algebra, science, writing, civics. Lee Webb et al. (2008) analyzed 24 items of the PPVT-III but only one item was biased and it favored blacks.

Several studies employed the Standardization P-DIF method. Schmitt & Dorans (1990) reported that the SAT verbal displayed few DIFs and that, among those DIFs items, both forms (3H and 4H) were biased in favor of mexicans compared to whites whereas the 3H form (and 4H form) displayed small bias against (and moderate bias in favor of) blacks. Freedle & Kostin (1997) found no evidence of bias against blacks in the SAT verbal and GRE verbal. Likewise, Scherbaum & Goldstein (2008) showed DIF cancellation in the analysis of the 11 items of the civics knowledge test.

Several studies used the Logistic Regression Method. Crane et al. (2004) analyzed the CASI, a test of cognitive functioning assessing a broad range of cognitive abilities, and reported only 2 biased items out of the 41 items of the test with respect to black-white groups and 2 biased items as well with respect to gender groups in a sample of elderly people. Meiring et al. (2005) analyzed an English reading comprehension test administered to black and white students in South Africa and reported DIFs of small magnitude. Hinton (2015) found no bias against blacks at the total score level both in the Raven APM and Raven SPM.

Although well designed and versatile, the SIBTEST was rarely used. Richwine (2009) compared the performance between US immigrants such as Mexicans and hispanics and native US whites on the PIAT-math but found no consistent bias against minorities when using SIBTEST.

Probably the most important DIF method is IRT, which was often employed. Linn et al. (1981) analyzed the reading comprehension test of the Metropolitan Achievement Tests in a

large sample of blacks and whites but reported only few DIFs which tended to cancel out. Drasgow (1987) also found small DIFs going in both directions in the ACT-Math and ACT-English, which eventually produced DIF cancellation between blacks and whites. Shepard et al. (1984) found small bias in a math achievement test against blacks on randomly equivalent and large samples of blacks and whites. Gibson (1998) found that the items of the ASVAB are riddled with DIFs but without consistent bias against minorities such as blacks or hispanics, as some subtests were strongly biased against whites and most often they were biased in favor of blacks. Scherbaum & Goldstein (2008) analyzed the 11 items of the civics knowledge test and found that 2 items were biased against blacks.

A more recent DIF method is the Latent Class Analysis (LCA) which examines latent group (or class) variable rather than manifest, observed group variable (e.g., race, gender). This is a cluster method similar to Factor Analysis but for finding and grouping patterns of item responses. After DIF items were detected with LCA, several analyses such as ANOVA or Logistic Regression are typically carried out to check whether manifest group variables correlate with latent classes. Not many studies used LCA for test bias but Lee Webb et al. (2008) reported that black-white race variable was not correlated with latent classes in the PPVT-III while Hinton (2015) reported no correlation between black-white race and latent classes in the Raven APM and Raven SPM.

DIF test of gender bias was also widely studied. Focusing only on IRT method, many studies reported acceptable comparability across gender groups for the Raven APM (Abad et al., 2004; Chiesi et al., 2012), the WISC-III (Maller, 2000), the Kaufman Adolescent and Adult Intelligence Test (Immekus & Maller, 2009), the vocabulary test of the Open Source Psychometrics Project (Kirkegaard, 2021).

When the test is composed of multiple subtests each consisting of multiple items, another method known as Multi-Group Confirmatory Factor Analysis (MGCFA) can be carried out to explore the comparability of means and covariances. The method proceeds by adding additional constraints to the initial, free model. The model parameters are iteratively constrained to be the same across groups. The following steps are taken: first, constrain the factor structure (configural invariance), second, constrain the factor loadings (weak/metric invariance), third, constrain the intercepts (strong/scalar invariance), fourth, constrain the residuals (strict invariance). If the model fit shows a decrement throughout one of the steps, group invariance (or equivalence) is rejected. Since the last equality constraint is sometimes dropped due to confoundings between measurement errors and specific variance (Dolan & Hamaker, 2001, p. 15), strict invariance is not a condition for establishing measurement invariance.

The great majority of these studies confirm the group equivalence between whites and blacks in the WISC-R (Dolan, 2000), the K-ABC (Dolan & Hamaker, 2001), the K-ABC II (Scheiber, 2015), the WAIS-IV (Frisby & Beaujean, 2015), three of the Woodcock-Johnson standardization samples, WJ-I, WJ-R, and WJ-III (Hu, 2017), the WJ-IV (Hajovsky & Chesnut, 2022)[1], a large test battery composed of 19 subtests in the VES data (Lasker et al., 2021), the ASVAB test of the NLSY79 data (Lasker et al. 2021) the ASVAB test of the

[1] It should be noted that Hajovsky & Chesnut (2022) controlled for parent education when conducting their analysis, which is an improper approach.

NLSY97 (Hu et al., 2019), the DAS-II test (Trundt et al., 2018), the NIH toolbox cognition battery (Lasker et al., 2019). The few exceptions come mostly from studies reporting bias in South African samples (Dolan et al., 2004; Lasker, 2021).

Although the different methods reach an agreement about test fairness, methodological issues are present. One major issue shared among these aforementioned studies is that they often do not use a purification procedure. Typically, this is achieved by estimating the item response in the free model and then selecting the items not flagged as DIF to be used in the matching total score, also called the anchor. The remaining items are subject to DIF estimation. Not removing DIF items in the matching total score variable causes inflated Type I errors (Finch, 2005; Shih et al., 2014; Fikis & Oshima, 2017). Two other major problems were highlighted by DeMars (2010). First issue, measurement errors cause imperfect total score matching which will then produce false positives in the detection of DIF whenever the groups differ in total score. DIF methods which do not estimate latent scores or correct for measurement errors (for instance by using the SIBTEST correction; DeMars, 2009) should not be used unless the test has very high reliability. Second issue, all of the methods mentioned earlier (with the exception of IRT and SIBTEST) do not account for the guessing parameter, which then produces biased estimates of the item probabilities whenever guessing is present and groups differ in total score.

Highlighting these points is particularly important in light of the present study, which analyzes the Wordsum vocabulary test of the General Social Survey (GSS). Because the test has a reliability of 0.63 among blacks and 0.71 among whites (Hu, 2017) and because one should expect guessing to be present in any cognitive or achievement test, the ideal method must account for measurement error and guessing.

2. Method

2.1. Sample

The data used in this study comes from the freely available General Social Survey. The study sample comprises respondents in the cumulative survey years of 1974-2021. The sample was restricted to adults aged between 18 and 67, because at late age, around 70, an age-related crystallized ability decline takes place. Based on the Wordsum mean scores in the GSS data, this decline begins at age 68.[2]

Each of the 10 Wordsum items was coded as 0 and 1 for incorrect and correct response, while nonresponse was coded as incorrect. A few respondents got all items wrong. Since this outcome is not informative, these few cases were removed prior to the analysis.

The race variable does not specifically ask whether the respondent is also of hispanic ethnicity. This question was only available from the year 2000 onwards. For the present study, a black-white variable is built and coded so as to exclude the hispanic ethnicity in all years for which it was available, yet one could ask whether the race variable is biased before

[2] The weighted means were similar in the unweighted and weighted results. Sampling weights were calculated as the interaction of the oversamp and wtssall variables, as explained in an earlier article (Hu, 2017).

the year 2000. According to the GSS codebook (Appendix B), the white category in the race variable (before the year 2000) includes Mexicans, Spaniards and Puerto Ricans “who appear to be white”. However, except for the years 1998, 2000 and 2002, the interviewers are requested to ask explicitly about the respondent’s race whenever they had any doubt. Furthermore, an earlier article analyzing the black-white wordsum score gap over time revealed no impact at all regardless of whether Mexico, Puerto Rico and other Spanish were included or not (Hu, 2017).

Our study sample includes 2,826 blacks and 15,186 whites, of which 8,051 are females and 7,135 are males. Missing data was handled using complete observation based on all variables under study (items, gender, race). The sampling weight variable has been considered initially because correct inference to the population is of utmost importance. Accounting for weights affects parameter estimates, standard errors and model fit. However, the weighted results of the initial IRT analysis were unreliable.[3] Therefore, the results presented below are not weighted.

2.2. Analysis

The IRT analysis, which has the desirable property of accounting for measurement error and guessing, is performed using *mirt* package (Chalmers, 2012). In principle, item responses suspected to be riddled with guessing should account for this parameter by estimating a 3PL model. In the present study, the 2PL results are reported for both the race analysis and gender analysis because the 3PL produced large standard errors. A possible reason is that the guessing parameter is poorly estimated (Hambleton et al., 1991, p. 112). Nonetheless, the results from the 3PL model are available in the supplementary file.

The *mirt* 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 as constrained across groups. To compare nested models, *mirt* reports the Chi-Squared test (χ^2), Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Sample-Size Adjusted BIC (SABIC), Hannan-Quinn (HQ) Criterion, and the Benjamini-Hochberg (BH) p-value. The BH p-value adjusts for the false discovery rate (false positive) of a traditional p-value that is equal to the alpha level of significance. Because significance testing is still not informative enough, DIF effect size measure is provided. Unlike previous effect sizes designed for IRT, the Differential Response Functioning (DRF) statistics proposed by Chalmers (2022) has the advantage of accounting for the sampling variability of the item parameters across ability levels. Two DRF statistics are available: the signed DRF (sDRF), which reflects the group difference after accounting for the DIF cancellation due to crossing DIF, and the square root of the squared DRF (dDRF), which calculates the absolute value of the group difference in terms of deviation (instead of variance). Because these statistics are in the same metric as the expected scale scores, the interpretation of the DRF is straightforward. For instance, the sDRF value of +0.1

[3] Sampling weight was initially applied, and while the impact on estimates was very small, all items showed equal fit indices when testing for DIF. Whatever the cause, this outcome cannot be trusted.

(or -0.1) is interpreted as a 10% higher probability of item response for the focal (or reference) group whereas the dDRF value of 0.1 is interpreted as a 10% absolute difference in probability.

3. Result

3.1. Racial group

One major assumption of IRT is unidimensionality, which merely requires a dominant factor (Hambleton et al., 1991, pp. 9-10). A parallel analysis based on polychoric correlation for dichotomous variables is employed, using *random.polychor.pa* package for R (Presaghi & Desimoni, 2019). Results from the Factor Analysis solution based on 10 factors displays eigenvalues of 3.83 and 0.78 for the first factor and the second factor, respectively, in the white sample and eigenvalues of 3.01 and 0.69 in the black sample. Based on the relative size of the first and second factors, we confirm the presence of one dominant factor.

The following step is to estimate the item equivalence across groups. Table 1 displays the result of the DIF screening procedure with model fit indices for each constrained model (one item being freely estimated at a time) compared to the baseline model in which all items have equal slopes and intercepts. Based on the p-value, all items are significant. However, it is known that significance test typically yields very high Type I error, especially with large sample sizes. An alternative is to flag for DIF index values that display relatively larger values. The advantage of this exploratory approach is that if all flagged items have large effect sizes, one could flag more items based on fit indices. Here in Table 1, the items Word E, F, G show a serious departure from other items.

Table 1. Fit Indices For Wordsum Items Based on 2 PL Model for Racial Group

Items	Model Fit Indices					
	AIC	SABIC	HQ	BIC	X^2	BH p
Word A	-33.25	-24.01	-28.12	-17.65	37.25	<.001
Word B	-26.90	-17.71	-21.82	-11.35	30.95	<.001
Word C	-16.32	-7.08	-11.20	-0.73	20.32	<.001
Word D	-32.96	-23.72	-27.83	-17.36	36.96	<.001
Word E	-279.69	-270.45	-.274.56	-264.09	283.69	<.001
Word F	-139.70	-130.46	-134.57	-124.10	143.70	<.001
Word G	-153.82	-144.58	-148.69	-138.22	157.82	<.001
Word H	-20.30	-11.06	-15.17	-4.70	24.30	<.001
Word I	-7.39	1.85	-2.26	8.21	11.39	.003
Word J	-24.89	-15.65	-19.76	-9.29	28.89	<.001

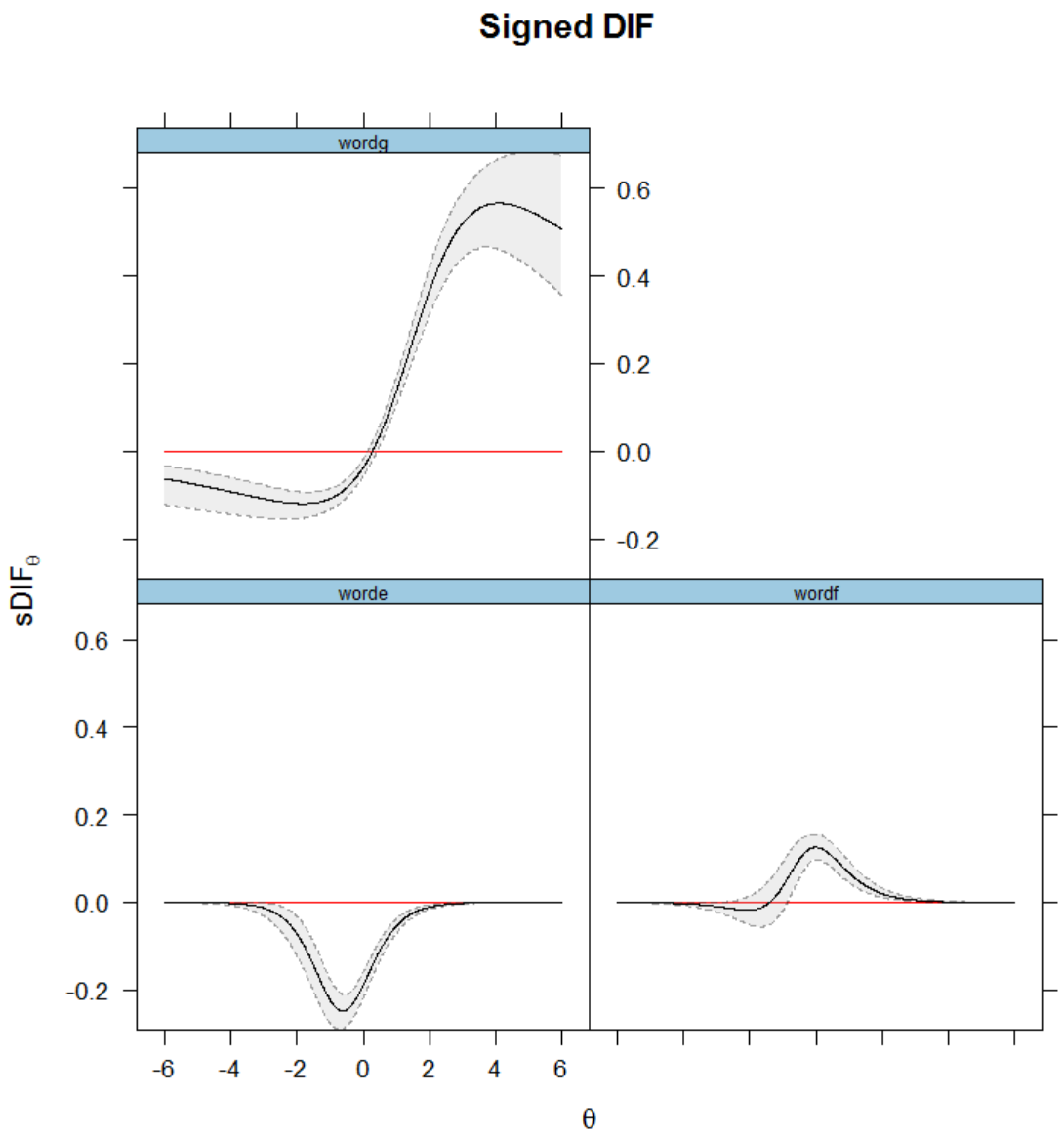
The next step is to estimate the multiple group IRT model by using as anchors the 7 other items not classified as DIF. Rather than relying on significance tests, the magnitude of DIF is estimated using Chalmer's DRF statistics.

The final step is to estimate the DIF sizes of the flagged items. Table 2 displays the signed DRF and dDRF statistics. With respect to the signed DRF, a positive (negative) value indicates an advantage for whites (blacks). The signed DRF statistics, the focus of this study, show that Word G is strongly biased against blacks (13% difference in probabilities) whereas Word E is moderately biased against whites (9,6% difference in probabilities). That Word F displays such a small signed DIF is important because it confirms that other items not selected as DIF would definitely not display DIF. Figure 1 plots the item probabilities given the DRF statistics. We observed that Word E is biased against whites at low and medium levels of ability whereas Word G is biased against blacks (whites) at high (low) levels of ability.

Table 2. DIF Effect Size based on the DRF Statistics for Racial Group

	Signed DRF (CI 95%)	dDRF (CI 95%)
Word E	-0.096 (-0.108;-0.083)	0.126 (0.110;0.142)
Word F	0.066 (0.050;0.083)	0.078 (0.060;0.097)
Word G	0.134 (0.108;0.159)	0.233 (0.199;0.267)

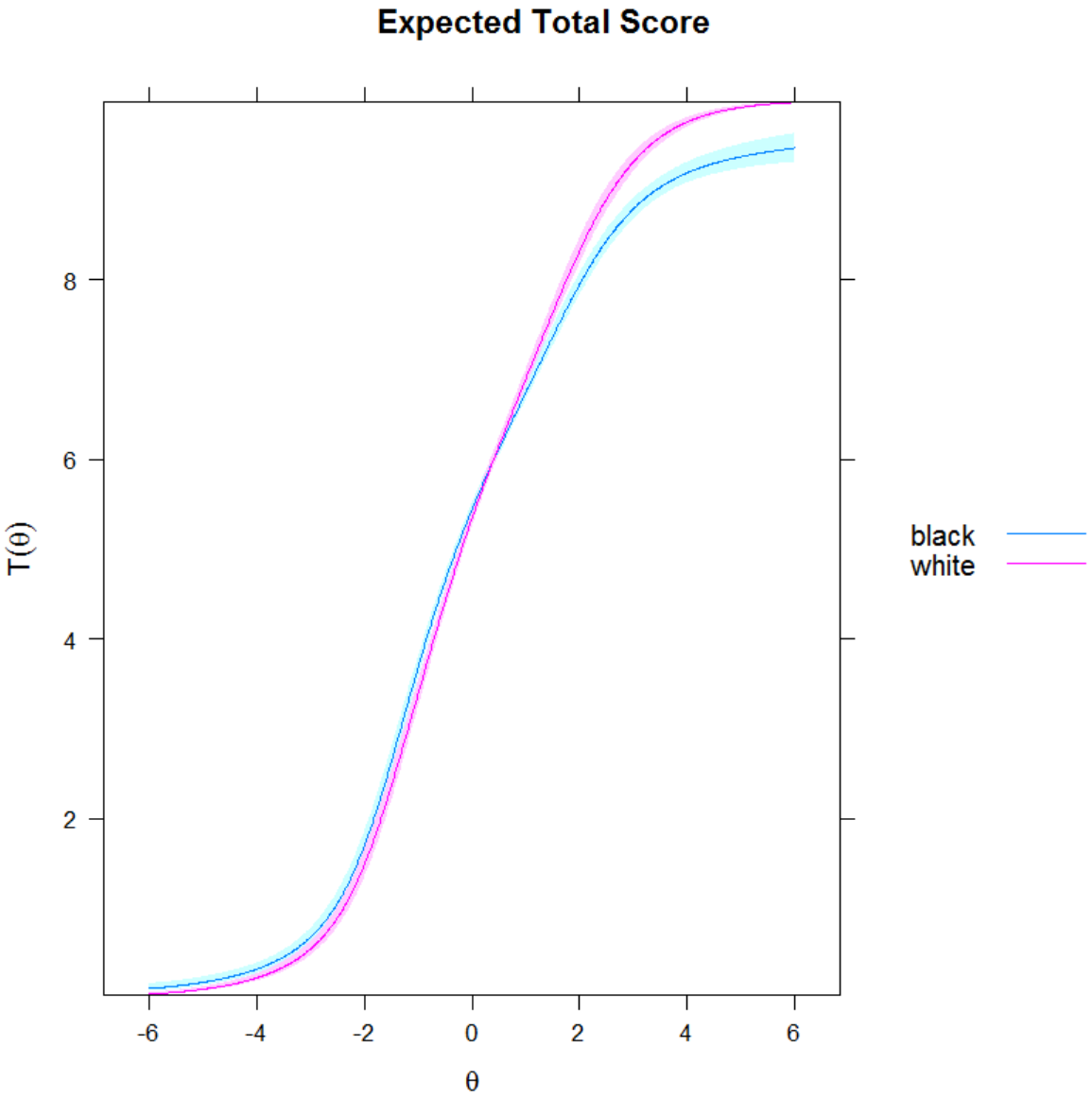
Figure 1. Differential Response Functioning Plots for Black and White Groups



Yet another crucial point is the impact of the combined DIFs at the test level. Figure 2 shows the latent score for each group. At low levels of latent score, blacks are slightly advantaged, while at very high level of latent score, whites are noticeably advantaged. This pattern mirrors the finding that the dDRF size is much larger than the signed DRF size, showing evidence of crossing DIF. But because this portion of the population is very small, the outcome is inconsequential.

To check the robustness of our analysis, the model is rerun by flagging one more item as DIF, based on fit index values. Word A and D are the next items having higher values. A first model, in which Word A, E, F, G are tested as DIFs, yields values of -0.046 for sDIF and 0.050 for dDIF. A second model, in which Word D, E, F, G are tested as DIFs, yields values of 0.013 for sDIF and 0.035 for dDIF. In both models, the sDIF and dDIF values for Word E, G and F are very similar to Table 2. The very low values of Word A and D therefore provides justification for treating these items as anchors.

Figure 2. Latent Total Score by Race



3.2. Gender group

The same analysis is conducted across gender in the white sample only since IRT typically requires large samples for ensuring stability of item parameters. As before, the unidimensionality assumption must be met. A parallel analysis based on polychoric correlation for dichotomous variables is employed. Results from the Factor Analysis solution based on 10 factors displays eigenvalues of 3.96 and 0.77 for the first factor and the second factor, respectively, in the male sample and eigenvalues of 3.69 and 0.82 in the female sample. Based on the relative size of the first and second factors, we confirm the presence of one dominant factor.

Table 2 displays the result of the DIF screening procedure with model fit indices for each constrained model (one item being freely estimated at a time) compared to the baseline

model in which all items have equal slopes and intercepts. Based on the p-value, nearly all items are significant. Because p values are extremely sensitive to large sample sizes, we can only ascertain that Word B, D and E are not displaying DIF while other items may just be false positives. Instead, an examination of the index values reveals that Word C, G, H and J display relatively larger values than other items. These 4 items will then be carefully examined.

Table 3. Fit Indices For Wordsum Items Based on 2 PL Model for Gender Group

Items	Model Fit Indices					
	AIC	SABIC	HQ	BIC	X^2	BH p
Word A	-12.74	-3.84	-7.68	2.52	16.74	<.001
Word B	3.93	12.83	8.99	19.19	0.07	.966
Word C	-102.26	-93.36	-97.20	-87.01	106.26	<.001
Word D	3.41	12.31	8.46	18.66	0.59	.825
Word E	1.34	10.24	6.39	16.59	2.66	.330
Word F	-9.01	-0.11	-3.95	6.25	13.01	.002
Word G	-39.96	-31.06	-34.90	-24.70	43.96	<.001
Word H	-64.80	-55.90	-59.74	-49.54	68.80	<.001
Word I	-3.95	4.95	1.11	11.30	7.95	.027
Word J	-39.06	-30.16	-34.00	-23.81	43.06	<.001

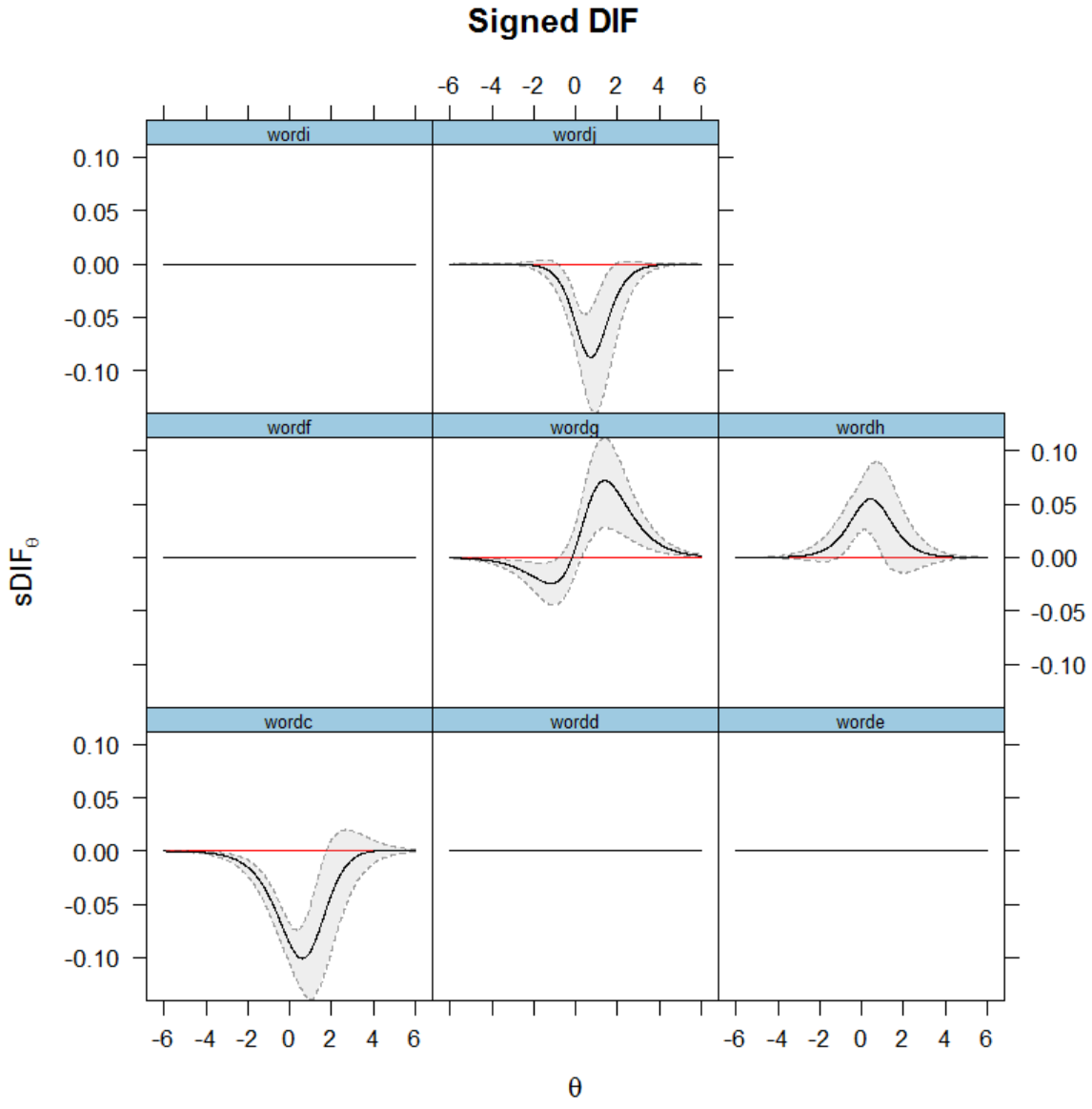
The next step is to estimate the multiple group IRT model by using as anchors the 6 other items not classified as DIF. As before, the magnitude of DIF is estimated using Chalmer's signed DRF and dDRF statistics. With respect to the signed DRF, a positive (negative) value indicates an advantage for females (males). Table 4 shows that the magnitude of DIF is very small for all items that were flagged as DIF. This means the other items selected for the anchor would definitely not display DIF. Figure 3 plots the item probabilities given the DRF statistics. That DIFs occur at the medium level of ability implies that the majority of the respondents is concerned but since the size is small and the DIFs cancel out, there is not much for concern. It is yet worth mentioning that Word G displays a complex crossing DIF, advantaging one group at some levels and the other group at other levels of ability. This pattern is reflected by the much larger value of the dDRF compared to the signed DRF statistics. This was true for the analysis of racial groups as well.

Table 4. DIF Effect Size based on the DRF Statistics for Gender Group

	Signed DRF (CI 95%)	dDRF (CI 95%)

Word C	-0.072 (-0.089;-0.053)	0.076 (0.056;0.097)
Word G	0.018 (-0.001;0.036)	0.039 (0.016;0.062)
Word H	0.038 (0.018;0.058)	0.041 (0.022;0.064)
Word J	-0.048 (-0.069;-0.027)	0.056 (0.031;0.084)

Figure 3. Differential Response Functioning Plots for Gender Group

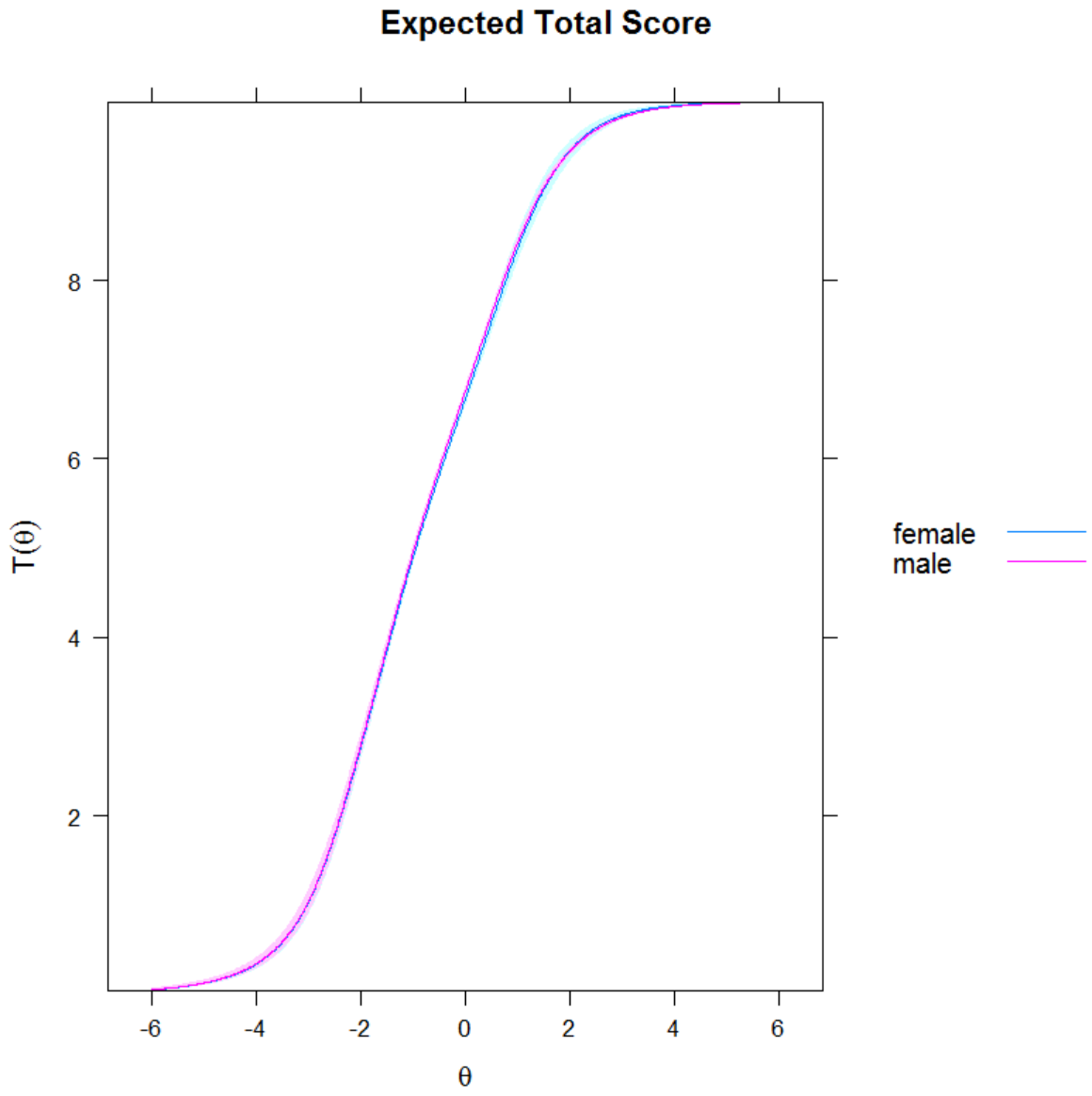


Once again, it is necessary to visualize the impact of the estimated DIFs at the total score level. Figure 4 shows that the total test curve is almost identical for both groups. The confidence bands around the curves are small, indicating accuracy in these estimates.

To check the robustness of our analysis, the model is rerun by considering Word G as part of the anchor, because Word G has the lowest sDIF/dDIF value. Such model which estimates

DIF values for only Word C, H and J produced values of sDIF and dDIF similar to Table 4. We do not attempt to rerun the model by flagging more items as DIFs because Word G already has very small effect size values, but also because the validity of internal test bias is called into question whenever DIFs are considered as a major phenomenon, as argued by DeMars & Lau (2011).

Figure 4. Latent Total Score by Gender



4. Discussion

In the present analysis, we established group equivalence with respect to item response probabilities for black-white group and gender group comparisons. Two items display large DIFs in the racial group analysis but the impact at the test level is inconsequential. Roznowski & Reith (1999) argued that DIFs can be retained if they do not cause poorer

measurement quality or deteriorate the test validity and provide sufficient information about individuals' performance.[4]

Other concerns are worth mentioning. One question regarding the anchor strategy used for purification pertains to the number of items used as matching criterion (i.e., total score). For the race group and gender group, only 7-item and 6-item anchors were used, respectively. Simulations from Wang & Yeh (2003) indicate that a 10-item anchor does not yield better results than a 4-item anchor as both performed well in terms of Type I error and power. Therefore, our anchor strategy yielded accurate results. Perhaps more concerning is that the Wordsum was relatively easy for whites (Hu, 2017, Figure 1). When the item difficulty closely matches the examinee ability, measurement precision increases and translates into reduced standard errors (Hambleton et al., 1991, p. 95). One other concern is the difference in test reliability between blacks (0.63) and whites (0.71) despite the IRT estimation of latent score which corrects for measurement errors. If reliability differences should introduce a bias, it will overestimate DIFs due to differences in matching. One last question, and the most problematic one, concerns the validity of internal test bias. A problem known as ipsitivity. This phenomenon takes the form of a pervasive bias affecting every item identically. Sound test construction and item reviews by content experts are known as valid practices for avoiding ipsitivity (Clauser & Mazor, 1998, pp. 286, 292; Penfield & Camilli, 2007, pp. 161-162). Because DIF techniques evaluate the relative size of DIF, the construction of the anchor must be ideally theory-driven rather than data-driven. Unfortunately, the Wordsum test items have not been thoroughly reviewed by experts like some renowned tests such as the Wechsler or the Kaufman's battery. For this reason, the present analysis is still informative but requires validation by content expertise.

It seems rather surprising that a culture loaded test such as the Wordsum displays only a small bias with respect to racial groups. Because culture bias only happens when two persons with equal total score are unequally exposed to the specific knowledge elicited by the item, one could infer that blacks and whites are equally exposed to the knowledge content used in the wordsum. Rowe et al. (1994, 1995) argued that equal exposure to general culture is totally expected due to the omnipresent force of the mass-market culture in developed countries. These authors established cultural equivalence between whites and blacks in large US samples.

Despite strong evidence of measurement equivalence, the validity of intelligence testing is still hotly debated and challenged. One such attempt comes from Fagan & Holland (2002). They designed a cognitive test using novel words, on which the black-white gap vanished in their small sample experimentation. This leads them to conclude that traditional IQ tests measure nothing else than knowledge. Not only inspection time tests (Pesta & Poznanski, 2008), backward digit span tests (Jensen & Figueroa, 1975; Dalliard, 2013), and measurement invariance cast doubt on their assumption, but their culture-fair test did not actually measure intelligence. Their students were trained on the meaning of these novel words (p. 365). The description is clearly that of a short-term memory test, which exhibits a small black-white gap but also a black advantage when IQ is held constant (Jensen & Reynolds, 1982; Reynolds & Jensen, 1983). And there is more. Not only their students were

[4] If one item is of any concern with respect to providing precise information about true ability, this would be Word G.

given several times the instruction to fill in everything through guessing, they were given several times the instruction to answer as quickly as possible (p. 367). The latter aspect bears similarity with the Coding (speeded) subtest of the ASVAB. But speeded subtests have the lowest correlation with g or reaction time (Jensen, 1980, p. 590, fn. 8; 1998, pp. 204, 224, 236; Coyle & Pillow, 2008, Fig. 2). Decades ago, Jensen (1980, ch. 14) reviewed similarly flawed attempts at designing a culture-free test to make the black-white gap vanish.

Research on test bias is not only 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., 2016). Another intriguing line of research pertains to the impact of cultural bias on heritability estimates of cognitive tests across groups (Pesta et al., 2020).

The GSS data offers a great variety of variables which can be examined along with the Wordsum test. Validating the cultural equivalence of the test implies that analyses involving a comparison between groups, such as the score difference over time (Hu, 2017), can now be justified. But as noted earlier, the low reliability of the Wordsum test requires the use of factor or latent score rather than the observed test score variable. Malhotra (2007) reported that the unique variance of each Wordsum item was not random but that a latent variable approach such as Structural Equation Model can effectively remove the distorting influence of unique variance. Still another issue is the ceiling effect in the white group. Methods such as tobit regression can estimate the latent score given floor or ceiling effect (Long, 1997). Although the Wordsum test needs improvement, it is still possible to accurately estimate the ability scores.

5. References

Abad, F. J., Colom, R., Rebollo, I., & Escorial, S. (2004). Sex differential item functioning in the Raven's Advanced Progressive Matrices: Evidence for bias. *Personality and individual differences, 36*(6), 1459-1470. https://doi.org/10.1016/s0191-8869(03)00241-1

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of statistical Software, 48*, 1-29. https://doi.org/10.18637/jss.v048.i06

Chalmers, R. P. S. (2016). A differential response functioning framework for understanding item, bundle, and test bias. *Unpublished doctoral dissertation, York University*.

Chalmers, R. P. (2022). A Unified Comparison of IRT‐Based Effect Sizes for DIF Investigations. *Journal of Educational Measurement*. https://doi.org/10.1111/jedm.12347

Chiesi, F., Ciancaleoni, M., Galli, S., Morsanyi, K., & Primi, C. (2012). Item response theory analysis and differential item functioning across age, gender and country of a short form of the advanced progressive matrices. *Learning and Individual Differences, 22*(3), 390-396. https://doi.org/10.1016/j.lindif.2011.12.007

Clauser, B. E., & Mazor, K. M. (1998). Using Statistical Procedures To Identify Differentially Functioning Test Items. An NCME Instructional Module. *Educational Measurement: issues and practice, 17*(1), 31-44.

Conoley, C. A. (2003). Differential item functioning in the Peabody Picture Vocabulary Test-: Partial correlation versus expert judgment. *Unpublished doctoral dissertation, Texas A&M University*.

Coyle, T. R., & Pillow, D. R. (2008). SAT and ACT predict college GPA after removing g. *Intelligence, 36*(6), 719-729. https://doi.org/10.1016/j.intell.2008.05.001
Crane, P. K., Belle, G. V., & Larson, E. B. (2004). Test bias in a cognitive test: differential item functioning in the CASI. *Statistics in Medicine, 23*(2), 241-256. https://doi.org/10.1002/sim.1713
Dalliard, M. (2013). Racial Differences on Digit Span Tests. *Human Varieties*. Retrieved from: https://humanvarieties.org/2013/12/21/racial-differences-on-digit-span-tests/
DeMars, C. E. (2009). Modification of the Mantel-Haenszel and logistic regression DIF procedures to incorporate the SIBTEST regression correction. *Journal of Educational and Behavioral Statistics, 34*(2), 149-170. https://doi.org/10.3102/1076998607313923
DeMars, C. E. (2010). Type I error inflation for detecting DIF in the presence of impact. *Educational and Psychological Measurement, 70*(6), 961-972. https://doi.org/10.1177/0013164410366691
DeMars, C. E., & Lau, A. (2011). Differential item functioning detection with latent classes: how accurately can we detect who is responding differentially?. *Educational and Psychological Measurement, 71*(4), 597-616.
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
Drasgow, F. (1987). Study of the measurement bias of two standardized psychological tests. *Journal of Applied psychology, 72*(1), 19. https://doi.org/10.1037/0021-9010.72.1.19
Fagan, J. F., & Holland, C. R. (2002). Equal opportunity and racial differences in IQ. *Intelligence, 30*(4), 361-387. https://doi.org/10.1016/s0160-2896(02)00080-6
Fikis, D. R., & Oshima, T. C. (2017). Effect of purification procedures on DIF analysis in IRTPRO. *Educational and Psychological Measurement, 77*(3), 415-428. https://doi.org/10.1177/0013164416645844
Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological Measurement, 29*(4), 278-295. https://doi.org/10.1177/0146621605275728
Freedle, R., & Kostin, I. (1997). Predicting Black and White differential item functioning in verbal analogy performance. *Intelligence, 24*(3), 417-444. https://doi.org/10.1016/s0160-2896(97)90058-1
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
Gibson, S. G. (1998). Gender and ethnicity-based differential item functioning on the Armed Services Vocational Aptitude Battery. *Unpublished master's thesis. Virginia Polytechnic Institute and State University, Blacksburg, VA*.
Hajovsky, D. B., & Chesnut, S. R. (2022). Examination of differential effects of cognitive abilities on reading and mathematics achievement across race and ethnicity: Evidence with the WJ IV. *Journal of School Psychology, 93*, 1-27. https://doi.org/10.1016/j.jsp.2022.05.001

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory (Vol. 2). *Sage*.
Hinton, D. (2015). *Uncovering the root cause of ethnic differences in ability testing: differential test functioning, test familiarity and trait optimism as explanations of ethnic group differences* (Doctoral dissertation, Aston University).
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
Immekus, J. C., & Maller, S. J. (2009). Item parameter invariance of the Kaufman Adolescent and Adult Intelligence Test across male and female samples. *Educational and Psychological Measurement, 69*(6), 994-1012. https://doi.org/10.1177/0013164409344489
Jensen, A. R. (1980). Bias in mental testing. *New York: Free Press*.
Jensen, A. R. (1998). The factor. *Westport, CT: Prager*.
Jensen, A. R., & Figueroa, R. A. (1975). Forward and backward digit span interaction with race and IQ: predictions from Jensen's theory. *Journal of Educational Psychology, 67*(6), 882. https://doi.org/10.1037/0022-0663.67.6.882
Jensen, A. R., & Reynolds, C. R. (1982). Race, social class and ability patterns on the WISC-R. *Personality and Individual Differences, 3*(4), 423-438. https://doi.org/10.1016/0191-8869(82)90007-1
Kirkegaard, E. O. W. (2021). An examination of the openpsychometrics.org vocabulary test. *OpenPsych*. https://doi.org/10.26775/op.2021.07.05
Lasker, J. (2021). Interpreting Cross-cultural Bias in Psychological Assessments: An Empirical Example. https://doi.org/10.31234/osf.io/zwb4c
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 Webb, M. Y., Cohen, A. S., & Schwanenflugel, P. J. (2008). Latent class analysis of differential item functioning on the Peabody Picture Vocabulary Test–III. *Educational and Psychological Measurement, 68*(2), 335-351. https://doi.org/10.1177/0013164407308474
Linn, R. L., Levine, M. V., Hastings, C. N., & Wardrop, J. L. (1981). Item bias in a test of reading comprehension. *Applied psychological measurement, 5*(2), 159-173. https://doi.org/10.1177/014662168100500202
Long, J. S. (1997). Regression Models for Categorical and Limited Dependent Variables. Advanced Quantitative Techniques in the Social Sciences Number 7. *Thousand Oaks, CA: Sage*.
Lord, F. M. (1976). A Study of Item Bias Using Characteristic Curve Theory.
Malhotra, N., Krosnick, J. A., & Haertel, E. (2007). The psychometric properties of the GSS Wordsum vocabulary test. *GSS Methodological Report, 11*, 1-63.
Maller, S. J. (2001). Differential item functioning in the WISC-III: Item parameters for boys and girls in the national standardization sample. *Educational and Psychological Measurement, 61*(5), 793-817. https://doi.org/10.1177/00131640121971527
Meiring, D., Van de Vijver, A. J. R., Rothmann, S., & Barrick, M. R. (2005). Construct, item and method bias of cognitive and personality tests in South Africa. *SA Journal of Industrial Psychology, 31*(1), 1-8. https://doi.org/10.4102/sajip.v31i1.182
Nandakumar, R., Glutting, J. J., & Oakland, T. (1993). Mantel-Haenszel methodology for detecting item bias: An introduction and example using the Guide to the Assessment of Test Session Behavior. *Journal of Psychoeducational Assessment, 11*(2), 108-119. https://doi.org/10.1177/073428299301100201

Penfield, R. D., & Camilli, G. (2007). Differential item functioning and item bias. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol. 26, pp. 125–167). Amsterdam, The Netherlands: North-Holland.
Pesta, B. J., & Poznanski, P. J. (2008). Black–white differences on IQ and grades: The mediating role of elementary cognitive tasks. *Intelligence, 36*(4), 323-329. https://doi.org/10.1016/j.intell.2007.07.004
Pesta, B. J., Kirkegaard, E. O., te Nijenhuis, J., Lasker, J., & Fuerst, J. G. (2020). Racial and ethnic group differences in the heritability of intelligence: A systematic review and meta-analysis. *Intelligence, 78*, 101408. https://doi.org/10.1016/j.intell.2019.101408
Presaghi, F., & Desimoni, M. (2019). random. polychor. pa: A parallel analysis with polychoric correlation matrices. *R package version, 1*, 4-3.
Reynolds, C. R., & Jensen, A. R. (1983). WISC-R subscale patterns of abilities of Blacks and Whites matched on Full Scale IQ. *Journal of Educational Psychology, 75*(2), 207. https://doi.org/10.1037/0022-0663.75.2.207
Reynolds, C. R., Willson, V. L., & Chatman, S. R. (1984). Item bias on the 1981 revision of the Peabody Picture Vocabulary Test using a new method of detecting bias. *Journal of Psychoeducational Assessment, 2*(3), 219-224. https://doi.org/10.1177/073428298400200306
Richwine, J. (2009). *IQ and immigration policy* (Doctoral dissertation, Harvard University).
Rowe, D. C., Vazsonyi, A. T., & Flannery, D. J. (1994). No more than skin deep: Ethnic and racial similarity in developmental process. *Psychological Review, 101*(3), 396. https://doi.org/10.1037/0033-295x.101.3.396
Rowe, D. C., Vazsonyi, A. T., & Flannery, D. J. (1995). Ethnic and racial similarity in developmental process: A study of academic achievement. *Psychological Science, 6*(1), 33-38. https://doi.org/10.1111/j.1467-9280.1995.tb00301.x
Roznowski, M., & Reith, J. (1999). Examining the measurement quality of tests containing differentially functioning items: Do biased items result in poor measurement?. *Educational and psychological Measurement, 59*(2), 248-269. https://doi.org/10.1177/00131649921969839
Scheiber, C. (2015) Do the Kaufman Tests of Cognitive Ability and Academic Achievement Display Ethnic Bias for Students in Grades 1 through 12. *Unpublished doctoral dissertation, Alliant International University*.
Scherbaum, C. A., & Goldstein, H. W. (2008). Examining the relationship between race-based differential item functioning and item difficulty. *Educational and Psychological Measurement, 68*(4), 537-553. https://doi.org/10.1177/0013164407310129
Scheuneman, J. D., & Gerritz, K. (1990). Using differential item functioning procedures to explore sources of item difficulty and group performance characteristics. *Journal of Educational Measurement, 27*(2), 109-131. https://doi.org/10.1111/j.1745-3984.1990.tb00737.x
Schmitt, A. P., & Dorans, N. J. (1990). Differential item functioning for minority examinees on the SAT. *Journal of Educational Measurement, 27*(1), 67-81. https://doi.org/10.1111/j.1745-3984.1990.tb00735.x
Shepard, L., Camilli, G., & Williams, D. M. (1984). Accounting for statistical artifacts in item bias research. *Journal of Educational Statistics, 9*(2), 93-128. https://doi.org/10.2307/1164716
Shih, C. L., Liu, T. H., & Wang, W. C. (2014). Controlling type I error rates in assessing DIF for logistic regression method combined with SIBTEST regression correction procedure and

DIF-free-then-DIF strategy. *Educational and Psychological Measurement, 74*(6), 1018-1048. https://doi.org/10.1177/0013164413520545
Stricker, L. J. (1982). Identifying test items that perform differentially in population subgroups: A partial correlation index. *Applied Psychological Measurement, 6*(3), 261-273. https://doi.org/10.1177/014662168200600302
te Nijenhuis, J., Willigers, D., Dragt, J., & van der Flier, H. (2016). 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
Trundt, K. M., Keith, T. Z., Caemmerer, J. M., & Smith, L. V. (2018). Testing for construct bias in the Differential Ability Scales: A comparison among African American, Asian, Hispanic, and Caucasian children. *Journal of Psychoeducational Assessment, 36*(7), 670-683. https://doi.org/10.1177/0734282917698303
Wang, W. C., & Yeh, Y. L. (2003). Effects of anchor item methods on differential item functioning detection with the likelihood ratio test. *Applied Psychological Measurement, 27*(6), 479-498. https://doi.org/10.1177/0146621603259902
Willson, V. L., Nolan, R. F., Reynolds, C. R., & Kamphaus, R. W. (1989). Race and gender effects on item functioning on the Kaufman Assessment Battery for Children. *Journal of School Psychology, 27*(3), 289-296. https://doi.org/10.1016/0022-4405(89)90043-5