Spearman's *g* Explains Black-White but not Sex Differences in Cognitive Abilities in the Project Talent

Meng Hu*
Independent researcher
* Contact: mh19870410@gmail.com

Abstract

The weak form of Spearman's Hypothesis, which states that the racial group differences are primarily due to differences in the general factor (*g*), was tested and confirmed in this analysis of the Project Talent data, based on 34 aptitude tests among 9th-12th grade students. Multi-Group Confirmatory Factor Analysis (MGCFA) detected small-modest bias with respect to race but strong bias with respect to within-race sex cognitive difference. After establishing partial measurement equivalence, SH was tested by comparing the model fit of correlated factors (non-*g*) model with a bifactor (*g*) model as well as the relative contribution of *g* factor means to that of the specific factors. While *g* was the main source of the Black-White differences, this wasn't the case for within-race sex differences. The average proportion of the score gaps accounted for by *g* is large (.73/.90) for the Black-White analysis but modest (.43/.50) for the sex analysis. The evidence of measurement bias in the sex analysis may cause ambiguity in interpreting SH for sex differences. Results from MGCFA were somewhat corroborated by the Method of Correlated Vectors, with high correlations of subtests' loadings with Black-White differences but near-zero correlations with sex differences. This finding replicates earlier MGCFA studies supporting SH with respect to the Black-White cognitive gap as well as earlier MGCFA studies revealing stronger gender bias than racial bias.

Keywords: Project Talent, Black-White IQ gap, Sex IQ gap, measurement invariance, MGCFA, MCV, Spearman's Hypothesis

1. Introduction

Large differences in cognitive abilities between U.S. race/ethnic groups, e.g., Blacks, Whites, and Hispanics, are beyond dispute (Murray, 2021). Jensen (1998, ch. 14) proposed that the magnitude of the racial differences in IQ, at least between Black and White Americans, as well as differences in cognitive-related socio-economic outcomes are a function of the *g*-loadings (i.e., the correlation between the tests or outcomes and the general factor of intelligence) of the respective cognitive tests and outcomes, making the *g* factor an essential element in the study of socio-economic inequalities. More specifically, Jensen (1985) proposed that race/ethnic group differences on cognitive tests are largely due to latent differences in general mental ability. This is known as Spearman's hypothesis (SH) which exists in two forms: the strong and the weak form, the latter of which was endorsed by Jensen. The strong form affirms that the differences are solely due to *g* factor differences while the weak form affirms that the differences are mainly due to differences in *g*. The alternative contra hypothesis states that group differences reside entirely or mainly in the tests' group or broad factors and/or test specificity and that *g* differences contribute little or nothing to the overall ones.

Unfortunately, Spearman's Hypothesis is not always well understood. Several researchers (e.g., Van der Sluis et al., 2006, 2008) misconstrued SH by interpreting a small *g* difference between groups as a rejection of SH. As Jensen (1998) mentioned, all tests measure *g* to some extent, some better than others. If the group differences in observed total IQ scores are small, as in the case of gender groups, one should not expect large *g* differences. The test of Spearman's *g* is about finding the proportion in the patterns of subtest score differences that is due to the general factor relative to non-*g* (e.g., specific) factors. Tests which better measure *g* would exhibit greater group differences. To answer this question, Jensen (1998) devised the Method of Correlated Vectors (MCV) to evaluate the magnitude of the subtest group differences that is due to *g*.

Prior to testing for SH, test score comparability must be established in order to produce unbiased estimates of means in specific and general factors, thus avoiding ambiguity in interpreting SH. This is best achieved through latent variable techniques at the item-level, such as Item Response Theory (IRT), and at the subtest-level, such as Multi-Group Confirmatory Factor Analysis (MGCFA).

The traditional view of culture bias holds that members of two groups, after being perfectly matched for latent ability, do not have equal probability of correct response on any given item. To achieve test comparability in MGCFA, members of different groups should use the same latent abilities (e.g., verbal, perceptual) to solve any given subtest, members of different groups should have equivalent subtest loadings (i.e., weights) on the latent factors, members of different groups matched in latent factor mean should get the same score on the subtests loading onto this latent factor. If the latent factor scores do not account fully for the group difference in the subtest scores, the remainder is due to external influence, commonly assumed to be culture bias. Group differences in subtest loadings and means are identified as metric and scalar non-invariance, respectively. Non-invariance typically comes from unwanted nuisance factors beyond the factor(s) that are the intended target of the measures. Millsap & Olivera-Aguilar (2012, p. 388) provides an illustration: the inclusion of a math test having a mixture of multiple-choice items and problem-solving items, with the latter being story problems, may introduce bias against foreign language speakers due to the verbal content of this test. If the math factor is not supposed to measure verbal skill, then such a test should be discarded.

Numerous studies have been conducted using MGCFA, mostly from US samples. There was a strong agreement that cognitive tests are cross-culturally valid between Whites and Blacks with minimal or no bias (Beaujean & McGlaughlin, 2014; Dolan, 2000; Dolan & Hamaker, 2001; Frisby & Beaujean, 2015; Hajovsky & Chesnut, 2022;[1] Hu et al., 2019; Kane & Oakland, 2010;[2] Keith et al., 1995; Lasker et al., 2019, 2021; Lubke et al., 2003; Scheiber, 2015, 2016a; Sipe, 2005; Trundt et al., 2018). There were two notable exceptions. One comes from Scheiber (2016b) who found strong measurement bias in the analysis of the WISC-V between 777 White males and 830 White females, 188 Black males and Black 221 females, and 308 Hispanic males and Hispanic 313 females. MGCFA was applied to all of

[1] Their analysis of the WJ-IV is faulty because parent education was controlled for. This has the consequence of attenuating any possible measurement bias.
[2] Their analysis of the WJ-III test reported fit index values with only 2 decimals but the ∆GFI and ∆TLI of 0.01 are already a sign of test bias. The impact of the bias is unknown but is unlikely to be large.

these six groups simultaneously. Muthén & Asparouhov (2014) showed that MGCFA is not practical for testing many groups (>2). Another comes from Benson et al. (2020) who analyzed the UNIT2 norming sample and found that scalar invariance was rejected not only for race (Whites, Blacks, Asians) and ethnicity (Hispanic) groups but also for gender groups. Metric invariance was also rejected (i.e., loadings are not exactly equal between groups) for age and gender groups, suggesting that the UNIT2 overall is somewhat biased with respect to any group.[3]

So far the evidence of strong measurement bias in race differences comes mainly from studies conducted in African countries. Dolan et al. (2004) compared the Junior Aptitude Test (JAT) scores of South African Black and White students and found that both metric (i.e., subtest loadings) and scalar (i.e., subtest means) invariance are violated.[4] Lasker (2021) re-analyzed Cockroft et al. (2015) and compared the WAIS-III scores of undergraduate South African students enrolled at an English medium University to undergraduate UK university students, and found that metric and scalar invariance are rejected. Warne (2023) compared the WISC-III scores of Kenyan Grade 8 students in Nairobi schools to the American norm and the WAIS-IV scores of Ghanaian students who showed English fluency at high school or university to the American norm. While measurement equivalence was established for Ghanaian students, it was rejected for Kenyan students.[5]

Unlike race/ethnic differences being the focus of criticisms with respect to cross-cultural comparability, sex differences in cognitive abilities are usually not the focus of these attacks. Yet research employing MGCFA showed mixed evidence of gender fairness. Some studies reported small or no measurement bias (Chen et al., 2015;[6] Chen et al., 2020; Dombrowski et al., 2021; Keith & Reynolds, 2018, p. 875; Irwing, 2012; Keith et al., 2011; Palejwala & Fine, 2015;[7] Pezzuti & Orsini, 2016; Reynolds et al., 2008; Rodríguez-Cancino & Concha-Salgado, 2023; van der Sluis et al., 2006) while others reported non-trivial bias, intercepts almost always being the common source of bias (Arribas-Aguila et al., 2019; Dolan et al., 2006; Johnson & Bouchard, 2007; Lemos et al., 2013; Pauls et al., 2020; Pezzuti et al., 2020; Saggino et al., 2014; Van der Sluis et al., 2008; Walter et al., 2021). Although not ideal, the percentage of subtest bias is usually not so severe to the point that comparability is impossible. The conclusion that cognitive tests are gender biased should also be tempered by the difficulty to achieve full measurement equivalence in survey scales using traditional MGCFA (Van De Schoot et al., 2015) and by the exhaustive list of studies examining measurement bias at the item-level, rather than subtest-level, showing only minimal bias against either race or gender groups (Hu, 2023). The lesson to be drawn is that a comprehensive study of test bias should employ both item-level analysis such as IRT and test-level analysis such as MGCFA.

---

[3] Their analysis however is faulty on multiple grounds. The analysis of age groups and gender groups was not disaggregated by race or ethnicity groups. Similarly, the analysis of ethnic groups (Hispanic vs. non-Hispanic) is confounded by race identity. The analysis of race groups was done using three groups instead of two.
[4] A limitation of their study is that MGCFA was applied to three groups simultaneously instead of two.
[5] Although Warne concluded that the Kenyan sample showed measurement equivalence, the ΔCFI was extremely high (.012) for scalar invariance.
[6] Despite their conclusion, scalar invariance is rejected on the basis of the large ∆RMSEA (.014). The abnormal change in $\chi 2$ is another red flag. CFI was reported with 2 decimals instead of 3, making it impossible to precisely evaluate ∆CFI. How many subtests' means have to be freed is unknown.
[7] They only use CFI and reported this value with 2 decimals instead of 3.

While measurement equivalence with respect to racial groups is well established in Western countries, only a few studies have tested the Spearman's Hypothesis (SH). So far, there have been two methods proposed for testing SH within MGCFA. Dolan (2000) proposed that the most parsimonious *g* model must fit better than the non-*g* model. Dolan et al. (2006) and, later, Frisby & Beaujean (2015) proposed that the group differences in *g* factor means cannot be fixed to zero in a *g* model without a serious worsening in model fit. Model comparison between a correlated-factors (non-*g* model) and a higher-order factor (*g* model) has been evaluated by Dolan (2000) and Dolan & Hamaker (2001) but these models fit almost equally well, although admittedly the bifactor model has not been tested and the contribution of *g* to the subtest difference is large. The constraint on *g* factor mean differences has been tested by Kane & Oakland (2010), Frisby & Beaujean (2015), Hu et al. (2019), Lasker et al. (2019;[8] 2021) mostly on a bifactor model and the results have been supportive of the Spearman's Hypothesis. A decomposition of the percentage of each subtest's difference due to *g* provides a clearer picture of the relevance of *g* versus specific factors. This strategy has not been commonly used, with the exception of Dolan (2000, Table 8). It requires multiplying the factor mean difference by the subtest's loadings. These numbers are often reported in the study of Black-White differences but not in the study of sex differences.

For this reason, the test of SH with respect to sex differences is much less conclusive. But this is also partly due to faulty methodologies. For instance, Van der Sluis et al. (2006, 2008) analyzed two twin samples from the Netherlands and another twin sample from Belgium. Not only they did not use adequate cutoffs for fit indices and merely report CFI and RMSEA (which is found in later studies to be very insensitive to misfit) with 2 decimals instead of 3, but they also entirely relied on tests of significance for testing the group difference in latent means. In the 2006 study of Dutch adult twins, it was found that one specific factor had a sex gap close enough to zero, while constraining the second-order *g* difference in means to zero will cause a misfit. This specification of *g* + a subset of first-order factors as best fitting model represents an alternative version of the weak SH model (Dolan, 2000) yet the authors concluded that SH was rejected and they did not even report the magnitude of the *g* difference or its contribution relative to specific factors. In the 2008 study of 12-13 years Dutch old twins and 9-13 years old Belgian twins, both data sets lacked power to reject either the strong SH (fixing all specific factor means to zero) or contra-SH model (fixing only the *g* factor means to zero), yet the Dutch and Belgian data yield *g* gaps of 3.83 and 1.58 IQ points, respectively, despite the specific factor means and loadings not being reported. Dolan et al. (2006) analyzed the WAIS-III in a subsample of the Spanish standardization data. After fitting a parsimonious weak SH model, they found a *g* gap close to zero and two specific factors showing non-trivial sex differences. Irwing (2012) analyzed the standardization sample of the WAIS-III using a bifactor model and reported Cohen's *d* gaps of .22 for *g*, but while the loadings were reported, the specific factor mean differences were not. Palejwala & Fine (2015) reported the *d* gaps of .21, .21, -.17 for *g*, Gs, and Gv, with Gsm factor fixed to zero, on the WPPSI-IV test, but the loadings are not reported. Reynolds et al. (2008) analyzed the gender differences in the KABC-II across different age groups between 6 and 18 years old. After fitting a parsimonious weak SH model, they discovered that the equality constraint on the *g* factor mean did not worsen model fit in 3 of the 4 age subgroups. In the

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

end, there is no compelling evidence that *g* is the main source of the subtest differences between sex groups.

So far SH has not been tested by directly comparing the correlated factors and the bifactor models. An advantage of the bifactor is that the specific and general factors are completely separated while the specific factors are represented as residuals in a higher order model (Bornovalova et al., 2020). This leads to several issues. The first is that the bifactor model produces purer measures of specific abilities (Murray & Johnson, 2013, p. 420). The second is that the higher order model posits, quite unrealistically, that the specific factors explain all the covariance among the observed test scores (Gignac, 2006a, 2006b), while the bifactor model posits that the specific factors account for the test scores' residual covariance that remains after extraction of the covariance that is due to *g* (Beaujean et al., 2014; Gignac, 2008). The third is that the proportionality constraints, imposed by the higher order but not by the bifactor, disallow any variation in the relative composition of variances attributable to specific abilities and *g* (Beaujean et al., 2014). Given that a more definite support of SH should involve partitioning the proportion due to the general factor and the proportion due to specific factors (Dolan, 2000, Table 8), it makes sense to take advantage of the bifactor structure.

While there are theoretical justifications for preferring a bifactor over a higher order factor structure, model comparison is complicated by the findings that fit indices used to evaluate models are biased in favor of the bifactor model when there are unmodelled complexities (Murray & Johnson, 2013). Yet a pro-bifactor bias is not a necessary outcome. Assuming no unmodeled misspecification, fit indices favor a correlated factors model when data were sampled from a true correlated factors structure, with unequal factor correlations (Morgan et al., 2015). When within-factor correlated residuals are misspecified, all fit indices correctly favor the correlated factors model regardless of conditions, except for SRMR, which incorrectly favors the bifactor model (Greene et al., 2019, Table 4). But whenever fit indices favor a bifactor structure, Murray & Johnson (2013) argued that the difference in fit must be very large to establish the superiority of the bifactor, in order to overcome this inherent bias. Given that the bifactor makes theoretical sense at explaining the structure of general intelligence, model comparison can still be made, while keeping in mind the aforementioned shortcomings.

2. Method

2.1. Data

The Project Talent is one of the largest studies ever conducted in the United States involving 377,016 9th-12th grade students during 1960 and drawn from all of the 50 states (Flanagan et al., 1962). The sample includes 4,481 twins and triplets from 2,233 families, and 84,000 siblings from 40,000 other families. The goal was to identify individuals' strengths (i.e., "talents") and steer them on to paths where those strengths would be best utilized. To this end, data on personal experiences, activities, aptitudes and abilities, health and plans for college, military service, marriage and careers were collected. Follow-up surveys were conducted until the students were age 29.

All analyses, including descriptive statistics, employ student weights, “BY_WTA”. The sample used in this study includes 70,776 White males, 71,381 White females, 2,443 Black males, 3,642 Black females with a weighted mean age of 15.9, 15.8, 16.0, and 15.8, respectively. The lower proportion of Black males compared to Black females may be explained by the higher likelihood of Black males dropping out of high school.

The Project Talent administered a considerable amount of tests, a great portion of which required specific knowledge. Detailed information provided by Wise et al. (1979). Major et al. (2012) considered the following 37 aptitude/cognitive subtests as cognitively relevant:

S1. Vocabulary (21 items). This scale gives some indication of the student’s general vocabulary.
S2. Literature (24 items). This scale measures familiarity with the world of literature, including prose and poetry.
S3. Music (13 items). This scale is intended to indicate the amount of musical information.
S4. Social Studies (24 items). This scale covers facts and concepts from the fields of history, economics, government and civics.
S5. Mathematics (23 items). This scale measures the vocabulary of mathematics, mathematical notation, and the understanding of mathematical concepts.
S6. Physical Science (18 items). This scale includes items about chemistry, physics, astronomy, and other physical sciences.
S7. Biological Science (11 items). This scale includes items about botany, zoology, and microbiology.
S8. Aeronautics and Space (10 items). This scale includes items about flying technique, navigation, jet planes, and space exploration.
S9. Electricity and Electronics (20 items). This scale stresses information that is acquirable through direct experience in the construction and maintenance of electrical and electronic equipment.
S10. Mechanics (19 items). This scale includes many items about automobiles and few others with common machines and tools related with mechanical activities.
S11. Art (12 items). This scale measures general knowledge about art, but excluding technical knowledge related to proficiency as an artist.
S12. Law (9 items). This scale measures general knowledge law that can be acquired through books or news reports concerning legal affairs.
S13. Health (9 items). This scale includes items related to practical health maintenance and nutrition, and common health care techniques.
S14. Bible (15 items). This scale measures general knowledge about the characters and teachings in the Bible.
S15. Theater (8 items). This scale has items dealing primarily with theater and ballet.
S16. Miscellaneous (10 items). This scale contains miscellaneous knowledge questions.
S17. Memory for Sentences (16 items). This scale measures the ability to memorize simple descriptive statements and to recall a missing word in a later sentence.
S18. Memory for Words (24 items). This scale measures another type of rote memory–the ability to memorize foreign words corresponding to common English words.
S19. Disguised Words (30 items). This scale measures the ability to form connections between letters and sounds.
S20. Word Spelling (16 items). This scale measures the ability to spell–not size of vocabulary.

S21. Capitalization (33 items). This scale indicates the degree of mastery of the rules of capitalization.
S22. Punctuation (27 items). This scale measures knowledge of the appropriate use of standard punctuation marks.
S23. English Usage (25 items). This scale measures knowledge of preferred usage.
S24. Effective Expression (12 items). This scale measures the ability to recognize whether an idea has been expressed clearly, concisely, and smoothly.
S25. Word Function in Sentences (24 items). This scale measures the sensitivity to grammatical structure.
S26. Reading Comprehension (48 items). This scale measures the ability to comprehend written materials, including passages on a wide range of topics.
S27. Creativity (20 items). This scale measures the ability to find ingenious solutions to a variety of practical problems.
S28. Mechanical Reasoning (20 items). This scale measures the ability to deduce the effects of the operation of everyday physical forces (e.g., gravity) and basic kinds of mechanisms (e.g., gears, pulleys, wheels, etc.)
S29. Visualization in 2D (24 items). This scale measures the ability to visualize how diagrams would look after being turned around on a flat surface in contrast to being turned over.
S30. Visualization in 3D (16 items). This scale measures the ability to visualize how a two dimensional figure would look after it had been folded to make a three-dimensional figure.
S31. Abstract Reasoning (15 items). This scale measures the ability to determine a logical relationship or progression among elements of a complex pattern.
S32. Arithmetic Reasoning (16 items). This scale measures the ability to reason in the manner required to solve arithmetic problems, but does not involve complex computation.
S33. High School Math (24 items). This scale measures mathematics taught up to 9th grade, and focuses mainly on elementary algebra, fractions, decimals, percents, square roots, intuitive geometry.
S34. Arithmetic Computation (72 items). This scale measures speed and accuracy of computation using the four basic operations and whole numbers.
S35. Table Reading (72 items). This scale measures speed and accuracy in a non-computational clerical task, involving obtaining information from tables.
S36. Clerical Checking (74 items). This scale measures speed and accuracy of perception in a simple clerical task, by determining whether the pairs of names are identical.
S37. Object Inspection (40 items). This scale measures speed and accuracy in perception of form, and requires to visually spot differences in small objects.

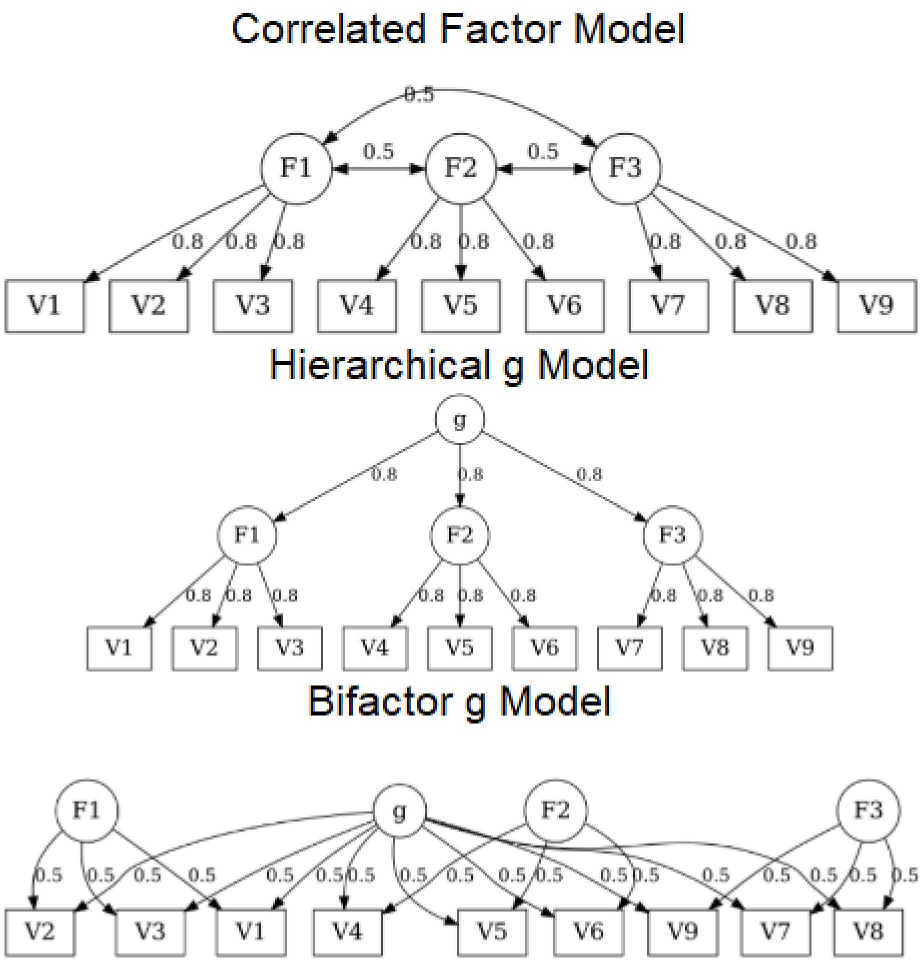
Three tests have been removed in the present analysis: memory for sentences (S17), memory for words (S18), and creativity (S27). The memory tests are highly correlated with each other but are poorly correlated with all other variables (between r=.10 and r=.20), which makes them unsuitable for CFA. Creativity has moderate correlations with other variables, has no main loading and its loadings are of modest or small size. Thus, a total of 34 aptitude/cognitive tests are used.

2.2. Analysis

All statistical analyses are done using R and, in particular, the *lavaan* package for MGCFA models. To test SH, competing models are employed, a correlated-factors (CF) as the non-*g* model, a higher-order factor (HOF) and a bifactor (BF) as representing two different structures of the *g* model. Another variation of the HOF structure is the Visual-Perceptual-Image Rotation (VPR) that was tested by Major et al. (2012) in Project Talent. In their study, the VPR-*g* model fitted much better than the CHC (Cattell-Horn-Carroll) HOF *g* model. The VPR was initially used in this study but it was found that the VPR model does not fit better than the CHC-based HOF model and produces sometimes inadmissible solutions such as negative variance. For this reason, the result for the VPR is not reported here but available in the supplementary material.[9] Figure 1 displays hypothetical competing CFA models that are investigated in the present analysis: 1) the correlated factors model which specifies that the first-order specific factors are correlated without the existence of a general factor, 2) the higher order factor model which specifies that the second-order general factor operates through the first-order specific factors and thus only indirectly influences the subtests, 3) the bifactor model which, unlike the higher order factor, specifies that both the general and specific factors have direct influences on the subtests.

Figure 1. Illustration of the competing CFA models

[9] Major et al. (2012) analyzed and used multiple imputation on the entire sample and separated the analysis by gender and by grade level (9-12). They included Memory for Words, Memory for Sentences, and Creativity subtests. In the present study, the VPR fits marginally better with a CFI=.002 at best, regardless of the subgroups being analyzed, and this remained true even after analyzing subgroups by grade level (9-12). In a supplementary analysis, available in the supplementary file, the VPR and CHC structures are compared following all the methodologies employed by Major et al. (2012), including adding back the two tests of memory and the test of creativity. The VPR model fits sometimes better in some grade levels, but in some other grade levels, the difference is marginal. This is a serious departure from Major et al. (2012) who found a large advantage for the VPR across all grade levels.



To evaluate and compare model specifications, fit indices such as CFI, RMSEA, $RMSEA_D$, SRMR and McDonald's Noncentrality Index (Mc) are used to assess model fit, along with the traditional $\chi2$. Higher values of CFI and Mc indicate better fit, while lower values of $\chi2$, RMSEA, $RMSEA_D$, SRMR indicate better fit. Simulation studies established the strength of these indices to detect misspecification (Chen, 2007; Cheung & Rensvold, 2002; Khojasteh & Lo, 2015; Meade et al., 2008). However, with respect to ΔRMSEA, doubts about its sensitivity to detect worse fit among nested models were raised quite often. Savalei et al. (2023) provided the best illustration of its shortcomings. According to them, this was expected because the initial Model A often has large degrees of freedom (df_A) relative to the degrees of freedom introduced by the constraints in Model B (df_B), resulting in very similar values of $RMSEA_B$ and $RMSEA_A$, hence a very small ΔRMSEA. For evaluating nested models, including constrained ones, their proposed $RMSEA_D$ solves this issue. $RMSEA_D$ is based on the same metric as RMSEA and is interpreted exactly the same way: a value of .08 suggests fair fit while a value of .10 suggests poor fit.

For overall model fit, Hu & Bentler (1999) recommended the following cutoffs based on a simulated 3-factor correlated model with 15 variables: a value close to .95 for CFI, .90 for Mc, .08 for SRMR, .06 for RMSEA would indicate good fit. This being noted, there is no such thing as a one-size-fits-all cutoff. Cheung & Rensvold (2001) explained that increased model complexity (e.g., increased number of indicators) has a tendency to reduce model fit. Sivo et al. (2006, Tables 8-10) found that the optimal cutoff value of fit indices for rejecting misspecified models depends on sample size: it decreases for Mc and increases for RMSEA.

A few studies have proposed fit index cutoffs for determining non-invariance. Meade et al. (2008) simulated multiple correlated factors models with varying levels of non-invariance and, assuming Type I error rate of .01, recommended a cutoff of .002 in ∆CFI to detect metric and scalar non-invariance while the cutoff for Mc depends on the number of factors and items (their Table 12), with most realistic conditions (i.e., up to 6 factors and up to 30 total items) lying between ∆Mc .0065 and .0120. Chen (2007) simulated a 1-factor model with varying the proportion of non-invariant indicators and pattern of non-invariance (unidirectional or bidirectional bias) and proposed several cutoffs: for testing loading invariance a change of ≥.005 in CFI, supplemented by a change of ≥.010 in RMSEA or a change of ≥.025 in SRMR; for testing intercept or residual invariance, a change of ≥.005 in CFI, supplemented by a change of ≥.010 in RMSEA or a change of ≥.005 in SRMR. The values of ∆Mc vary greatly depending on the condition and invariance steps (see Tables 4-6) but often lie between .010 and .015. Khojasteh & Lo (2015, Table 1) investigated the performance of fit indices in bifactor models for metric invariance and recommended the cutoffs .077-.101 for ΔMc, .003-.004 for ΔCFI, .021-.030 for ΔSRMR, .030-.034 for ΔRMSEA; with cutoffs smaller as sample sizes grow (from 400 to 1,200). These cutoffs will be considered together to evaluate model fit in the present study.

Sometimes, invariance does not hold. An interesting strategy is to compute the effect size to determine their importance. Gunn's et al. (2020) propose a standardized effect size called SDI, Signed Difference in expected Indicator, which provides the magnitude as well as the direction of the bias in standardized units similar to Cohen's *d*. A glaring issue is its dependence on the size of the observed SD of the "offending" subtest in the focal group. Groskurth (2023) proposes the Measurement Invariance Violation Indices (MIVIs) as effect sizes which are computed using the pooled SD of the latent factor. The latent SD has the advantage of being the same for all subtests loading onto that factor and consisting of true score variance only. Since observed SDs vary across subtests, the effect sizes are not comparable across subtests. At the same time, MIVIs are partially but not fully standardized due to not using the observed SD, making them comparable within but not across factors. MIVIs should yet produce more accurate effect sizes. These effect sizes have limited applications due to the assumption of invariant loadings when computing intercept differences or, more generally, the assumption of no cross loadings at all. Since effect sizes are still very useful, they will be computed whenever possible, having in mind these limitations.

MGCFA starts by adding additional constraints to the initial configural model, with the following incremental steps: metric, scalar, strict. A rejection of configural invariance implies that the groups use different latent abilities to solve the same set of item variables. A rejection in metric (loading) invariance implies that the indicators of a latent factor are unequally weighted across groups. A rejection in scalar (intercept) invariance implies that the subtest scores differ across groups when their latent factor means is equalized. A rejection in strict (residual) invariance implies there is a group difference in specific variance and/or measurement error. When invariance is rejected, partial invariance must release parameters until acceptable fit is achieved and these free parameters must be carried on in the next levels of MGCFA models. The variances of the latent factors are then constrained to be equal across groups to examine whether the groups use the same range of abilities to answer the subtests. The final step is to determine which latent factors can have their mean differences constrained to zero without deteriorating the model fit: a worsening of the model fit indicates that the factor is needed to account for the group differences. These model specifications will be presented in Table 1 further below.

While it is well established that measurement invariance requires that factor patterns, factor loadings and intercepts should be equal across groups. But there is no such agreement regarding residuals, which are composed of specific and error variances.

Several authors recommend strict invariance. Lubke & Dolan (2003) reported that a model with free residuals overestimates slightly the latent mean differences whenever the groups differ in their residuals because the model has to compensate for the differences in residuals. DeShon (2004, p. 146) explained that the common view that item specific variance is removed from the latent variable is based on the assumption that item uniquenesses are uncorrelated with each other or the latent variable. Violating this assumption will affect the estimation of the latent variables. Widaman & Reise (1997) argued that strict invariance has the advantage of having fewer parameters to estimate but the step can be skipped if the difference in error variance is justified (e.g., growth model with an age-related variable).

But other authors do not recommend strict invariance. Vandenberg & Lance (2000, p. 57) highlight the idea that latent variables are theoretically perfectly reliable, which makes strict invariance useless when evaluating latent means but useful when evaluating the reliability differences between groups. Little (1997, p. 55; 2013, p. 143) noted that strict invariance has a biasing effect if the group difference in residuals is small. Specifically, if the sum of the specific and random variance is not equal across groups, the amount of misfit that the constraints on the residuals would create must permeate all other estimated parameters.

Because the present analysis compares the contribution of each latent mean differences and model fit between competing models, strict invariance is ignored as it does not seem crucial for testing SH.

Table 1 presents a summary of possible models (including strict invariance levels that are ignored in the present study) for testing invariance and then *g*-models. The configural model allows group differences in loadings ($\lambda_1 \neq \lambda_2$), covariance matrix ($\Psi_1 \neq \Psi_2$), intercepts ($v_1 \neq v_2$), residuals ($\Theta_1 \neq \Theta_2$) and finally latent means equal to zero ($\delta=0$). The metric model adds group equality on loadings, then the scalar model adds group equality on subtests' means (i.e., intercepts), then the strict model adds group equality on the subtests' residuals (composed of specific and random variances). Only after scalar (or partial scalar) is set, that the latent factor means will differ across groups ($\delta \neq 0$). It is assumed that full invariance does not hold at all levels. In this case, the partial invariance at one level is carried on in the next models. Scalar (M3) and partial scalar (M3a) models will then be nested under M2a but not M2. Similarly, M4 and M4a are nested under M3a but not M3. Then, M5 adds a group equality on latent variances ($\Psi^*_1=\Psi^*_2$) and is nested under M4a. M6a specifies all non-*g* factor means to be zero, M6b specifies some non-*g* factor means to be zero, M6c specifies the *g* factor means to be zero. Understanding the nesting levels is important for the interpretation of $RMSEA_D$. For example, since M6a, M6b and M6c are competing models, all nested under M5, the $RMSEA_D$ for these models expresses their fit only with respect to M5, but not with respect to each other. The same principle applies to partial metric, partial scalar and partial strict. The $RMSEA_D$ expresses the fit of the partial model with respect to the previous level (M4a vs M3a, but not M4a vs M4).

Table 1. Summary of a typical MGCFA model

Model	Specification	Nesting
M1. Configural	$\lambda_1 \neq \lambda_2 + \Psi_1 \neq \Psi_2 + v_1 \neq v_2 + \Theta_1 \neq \Theta_2 + \delta=0$	
M2. Metric	M1 but adds (all) $\lambda_1=\lambda_2$	under M1
M2a. Partial Metric	M1 but adds (partial) $\lambda_1=\lambda_2$	under M1
M3. Scalar	M2a but adds (all) $v_1=v_2$ + (all)$\delta \neq 0$	under M2a
M3a. Partial Scalar	M2a but adds (partial) $v_1=v_2$ + (all)$\delta \neq 0$	under M2a
M4. Strict	M3a but adds (all) $\Theta_1=\Theta_2$	under M3a
M4a. Partial Strict	M3a but adds (partial) $\Theta_1=\Theta_2$	under M3a
M5. Lv variance	M4a but adds (all) $\Psi^*_1=\Psi^*_2$	under M4a
M6a. Strong SH	M5 but adds (all) $\delta_{non\text{-}g}=0$	under M5
M6b. Weak SH	M5 but adds (partial) $\delta_{non\text{-}g}=0$	under M5
M6c. No SH	M5 but adds $\delta_g=0$	under M5

The biggest concern regarding the validity of the results is the possible lack of power for the Black-White analysis due to the large sample unbalances, with a ratio of 1:31 and 1:21 for Black-White males and Black-White females, respectively. Yoon & Lai (2018) reported that a higher ratio of largest/smallest sample not only reduces power in detecting invariance but also decreases power in detecting misspecified models, as the fit indices show improvement as the sample unbalances increase. These authors proposed a subsampling approach. Consequently, slice_sample() from the *dplyr* R package was used to produce equal samples of Blacks and Whites, yielding a random sample of 2443 Whites for the male group and a random sample of 3642 Whites for the female group. Analyses were re-run 10 times using 10 random samples of White students in each gender group.

3. Result

3.1 Preparing data and testing assumptions

Missing data is handled with multiple imputation using *mice* package. Because the Predictive Mean Matching (PMM) method of imputation calculates the predicted value of target variable Y according to the specified imputation model, the imputation was conducted within race and within gender groups, totaling four imputations. It is inappropriate to impute the entire sample because it implies that the correlation pattern is identical across groups, an assumption that may not be true and may eventually conceal measurement non-invariance. The imputation is done for each subgroup conditioning that each case has at least 10 non-missing values (i.e., if a student only completed a few subtests, he/she was removed from the data prior to imputation). This ensures that only the students who provide enough information are used in the PMM method.

Maximum Likelihood (ML), used as the estimation method for CFA models, typically assumes normal distribution. Histograms show that the following subtests have a non-normal distribution: Math, Aeronautics, Electricity & Electronics, Capitalization, Word Functions, Table Reading. These variables are normalized, because achieving univariate normality helps achieving multivariate normality. All subtests variables are then z-score transformed because some variables vary so widely in their standard deviation, after normalization with power or log transformation, that it causes estimation problems.[10]

Univariate normality is then scrutinized. Curran et al. (1996) determined that univariate skewness of 2.0 and kurtosis of 7.0 are suspect, and that ML is robust to modest deviation from multivariate non-normality but that ML $\chi 2$ is inflated otherwise. Values for univariate kurtosis and skewness were acceptable, although the kurtosis values for Table Reading are a little high among White males (3.2) and White females (4.58). On the other hand, multivariate normality was often rejected. The multivariate non-normality displayed by the QQ plot was moderate for Black-White analysis in both male and female groups and sex analysis in the White group but perfectly fine for sex analysis in the Black group.

Exploratory Factor Analysis (EFA) was used to determine the appropriate number of factors. Similar to Major et al. (2012), it was found here that the 6-factor model was the most interpretable in all subgroups tested. The 4- and 5-factor models blend indicators into factors which are more ambiguous (math and english tests form one common factor; information and science tests form one common factor) and cause severe unbalances in factor sizes. The 7- and 8-factor models produce additional factors which are not associated with any particular ability or do not have any indicators with high loading. EFA reveals a large number of medium-size cross loadings. Since the results from simulation studies (Cao & Liang, 2023; Hsu et al., 2014; Xiao et al., 2019; Ximénez et al., 2022; Zhang et al., 2023) indicated that ignoring small cross loadings, typically set at .15 or .20 in these studies, has a tendency to reduce the sensitivity in commonly used fit indices, cross loadings are allowed when the average of the two groups is close to .20 but with a minimum of .15 per group.

The 6 factors in the best EFA model can be defined as English, Math, Speed, Information (or Knowledge), Science, and Spatial. From the perspective of the CHC structure, according to Major et al. (2012), they can be interpreted as, respectively, Reading & Writing Ability (Grw), Quantitative knowledge (Gq), Processing Speed (Gs), Comprehension-Knowledge (Gc), Science Knowledge (GK), and Visual Processing (Gv). In reality though, science knowledge has no counterpart in the CHC structure.

[10] It is indeed commonly suggested in statistics textbooks that researchers keep the original metric of the variables, but z-score transformation of all variables as opposed to only a few ones does not alter model fit or parameter estimates (at least, the fully standardized estimates) or even the standard errors. An advantage of z-score transforming all variables at once is to ease interpretation, since they are now on the same scale.

Initially, the MGCFA models were conducted without disaggregating by gender. But because it was found that the tests were biased with respect to gender, disaggregating by gender group appeared to be a more appropriate approach.

Finally, all analyses apply an equality constraint on the regression of each subtests on age. This is an important step because a non-invariance in these regressions implies that the effect of age on subtests differs between groups, which complicates group comparison.

3.2 Black-White analysis

Overall fit is acceptable in all models, except maybe for Mc. The configural and regression invariance both hold perfectly, thus only the next steps will be critically analyzed. Yet, due to potential lack of power, it was decided to investigate both metric and scalar levels by using modification indices to reveal the source of misfit based on the higher $\chi 2$ values, supplied by effect sizes whenever possible. Regarding strict invariance, this level is always severely rejected in the female group (ΔCFI=.005-.006) and rejected to a smaller extent in the male group (ΔCFI=.003). Details of these analyses are provided in the supplementary material.

Competing models are evaluated based on their optimal constraints. In the male group, the BF fits marginally better than the CF model, while the CF model fits much better than HOF. Given the fit indices being pro-bifactor biased, this superiority of the BF model is ambiguous. In the female group, BF fits much better than the CF model, but the HOF also fits much worse than the CF model. This finding of a worse fit for HOF is puzzling, especially for establishing the superiority of *g* models. Admittedly though, the BF model makes theoretically more sense than the HOF.

3.2.1 Black-White male group

The model specification is displayed as follows:

```
english =~ S1 + S19 + S20 + S21 + S22 + S23 + S24 + S25 + S26 + S31 + S34
math =~ S5 + S6 + S25 + S32 + S33 + S34
speed =~ S19 + S29 + S34 + S35 + S36 + S37
info =~ S1 + S2 + S3 + S4 + S5 + S6 + S7 + S8 + S11 + S12 + S13 + S14 +
S15 + S16 + S19 + S26
science =~ S1 + S6 + S7 + S8 + S9 + S10 + S28
spatial =~ S28 + S29 + S30 + S31 + S37
```

Table 2 contains a summary of the fit indices of the CF model and the free parameters. Using conventional cutoff criteria, and even strict criteria, all model constraints do not cause a serious deterioration in fit. At the metric level, three loadings display relatively larger $\chi 2$ values in modification indices. Partial metric removes the equality constraints and reveals non-trivial group differences in those loadings (based on their unstandardized units), despite no improvement in fit except for Mc. Effect sizes are not computed due to cross loadings. At the intercept level, two subtests display much larger $\chi 2$ values than other subtests in the modification indices. Partial scalar removes the constraints but it barely improves model fit, despite effect sizes being modest for Physical Science when using SDI (*d*=.36) and MIVIs (*d*=.31) and for Arithmetic Reasoning when using SDI (*d*=.35) and MIVIs (*d*=.33). These subtests are biased against Blacks. Adding first an equality constraint on latent covariances (M5) and then on latent variances (M6) does not worsen the fit in either case. A more parsimonious model (M7) adds a constraint on the speed factor mean due to having small group difference, but does not worsen the model fit.

Table 2. Black-White differences among males using Correlated Factors

Model Level	χ2	df	CFI	RMSEA	SRMR	Mc	RMSEAD [CI]
M0. Baseline	70578	495	.956	.044	.034	.620	
M1. Configural	69791	990	.955	.044	.033	.625	
M2. Regression	70088	1024	.955	.043	.034	.624	.014 [.012:.015]
M3. Metric	71549	1069	.954	.042	.035	.618	.028 [.026:.029]
M3a. Partial Metric[1]	71072	1066	.954	.042	.035	.620	.023 [.022:.025]
M4. Scalar	75161	1094	.951	.043	.035	.603	.059 [.058:.061]
M4a. Partial Scalar[2]	73762	1091	.952	.043	.035	.609	.051 [.049:.053]
M5. Lv covariance	74582	1106	.952	.043	.036	.605	.035 [.033:.037]
M6. Lv var-covariance	74827	1112	.951	.043	.036	.604	.032 [.028:.035]
M7. Lv reduced	**74962**	**1113**	**.951**	**.043**	**.036**	**.604**	**.047 [.039:.056]**

[1] Freed parameters (by descending order of χ2 size) are: English =~ Word Functions, Math =~ Word Functions, Speed =~ Disguised Words.

[2] Freed parameters (by descending order of χ2 size) are: Mechanics~1, Physical Science~1, Arithmetic Reasoning~1.

Note: higher values of CFI and Mc indicate better fit, while lower values of χ2, RMSEA, RMSEAD, SRMR indicate better fit.

Table 3 contains a summary of the fit indices of the HOF model and the free parameters. The appearance of good fit is once again misleading. At the metric level, three loadings display relatively larger χ2 values in modification indices. Partial metric releases them and this improves model fit. At the scalar level, Physical Science and Mechanics are once more associated with very large χ2 and released. The fit at the partial scalar barely improved. Constraining the latent variances (M5) to be equal does not worsen the model fit. This model can be taken as a more expanded version of the weak SH model because all factor means are estimated. Upon examining the factor means, English, Speed, Information display group differences close to zero. Their equality constraints do not worsen model fit (M6b). Compared to either models estimating *g* and non-*g* factor means, the strong SH (M6a) model fits barely worse (although the RMSEAD is close to .08, indicating not so good fit). It is probably safe to conclude that the weak SH model is superior.

Table 3. Black-White differences among males using Higher Order Factor

Model Level	χ2	df	CFI	RMSEA	SRMR	Mc	RMSEAD [CI]
M0. Baseline	84254	504	.947	.048	.040	.564	
M1. Configural	83357	1008	.946	.047	.040	.570	
M2. Regression	83657	1042	.946	.047	.040	.569	.014 [.012:.016]
M3. Metric	85893	1092	.944	.046	.042	.560	.033 [.031:.034]
M3a. Partial Metric[1]	84945	1089	.945	.046	.041	.564	.025 [.024:.027]
M4. Scalar	88575	1116	.942	.046	.041	.550	.056 [.054:.057]
M4a. Partial Scalar[2]	87413	1114	.943	.046	.042	.555	.048 [.046:.049]
M5. Lv variance	87615	1121	.943	.046	.042	.554	.025 [.022:.029]
M6a. Strong SH	88785	1127	.942	.046	.042	.550	.076 [.072:.079]
M6b. Weak SH	**87648**	**1124**	**.943**	**.046**	**.042**	**.554**	**.019 [.014:.024]**

[1] Freed parameters (by descending order of χ2 size) are: *g* =~ Math, English =~ Word Functions, Math =~ High School Math.

[2] Freed parameters (by descending order of χ2 size) are: Physical Science~1, Mechanics~1.
Note: higher values of CFI and Mc indicate better fit, while lower values of χ2, RMSEA, RMSEAD, SRMR indicate better fit.

Table 4 contains a summary of the fit indices of the BF model and the free parameters. Modification indices are used to detect the source of misfit. At the metric level, two loadings show much larger χ2. Partial metric shows small improvement. At the scalar level, the fit deteriorates very little, but the same subtests were the source of misfit; namely, Physical Science and Mechanics. Partial scalar barely improves model fit. Adding constraints on the latent variances (M5) does not affect model fit. This model, like its HOF counterpart, can be taken as a less parsimonious version of the weak SH. This model fits a little better than the strong SH (especially judging by RMSEAD) but clearly better than the no SH (ΔCFI=.003). A more parsimonious weak SH model (M6b) adds a constraint on english, speed and information factor means without worsening the model fit.

Table 4. Black-White differences among males using Bifactor

Model Level	χ2	df	CFI	RMSEA	SRMR	Mc	RMSEAD [CI]
M0. Baseline	65713	476	.959	.043	.028	.640	
M1. Configural	65836	952	.957	.043	.029	.642	
M2. Regression	66133	986	.957	.042	.030	.641	.014 [.012:.015]

M3. Metric	68480	1064	.956	.042	.032	.631	.026 [.025:.027]
M3a. Partial Metric[1]	67753	1062	.956	.041	.031	.634	.022 [.021:.023]
M4. Scalar	70835	1089	.954	.042	.032	.621	.053 [.051:.055]
M4a. Partial Scalar[2]	69688	1087	.955	.042	.032	.626	.043 [.041:.045]
M5. Lv variance	69961	1094	.955	.041	.032	.625	.029 [.026:.032]
M6a. Strong SH	71036	1100	.954	.042	.033	.620	.065 [.062:.069]
M6b. Weak SH	**69988**	**1097**	**.955**	**.041**	**.032**	**.625**	**.014 [.009:.019]**
M6c. No SH	74348	1098	.952	.043	.043	.606	.161 [.157:.166]

[1] Freed parameters (by descending order of χ2 size) are: *g* =~ High School Math, *g* =~ Word Functions.
[2] Freed parameters (by descending order of χ2 size) are: Physical Science~1, Mechanics~1.
Note: higher values of CFI and Mc indicate better fit, while lower values of χ2, RMSEA, RMSEAD, SRMR indicate better fit.

3.2.2 Black-White female group

The model specification is displayed as follows:

```
english =~ S1 + S13 + S19 + S20 + S21 + S22 + S23 + S24 + S25 + S26 + S31
+ S34
math =~ S5 + S25 + S32 + S33 + S34
speed =~ S19 + S34 + S35 + S36 + S37
info =~ S1 + S2 + S3 + S4 + S7 + S8 + S11 + S12 + S13 + S14 + S15 + S16 +
S19 + S26
science =~ S1 + S6 + S7 + S8 + S9 + S10
spatial =~ S28 + S29 + S30 + S31 + S37
```

Table 5 contains a summary of the fit indices of the CF model and the free parameters. Unlike the scenario in the male group, the female group usually displays larger fit decrements. At the metric level, modification indices reveal four loadings with larger χ2. Partial metric releases them but it barely improves fit. At the scalar level, values of ∆CFI=.004 and RMSEAD=.070 (which is close to .08) are concerning. Partial scalar releases Mechanics, Clerical Checking, Arithmetic Computation and Arithmetic Reasoning due to having by far the largest χ2. Model fit improves, especially for CFI, RMSEAD, and Mc. The effect sizes are nowhere small for Mechanics, using SDI (*d*=.67) and MIVIs (*d*=.48), and for Clerical Checking, using SDI (*d*=-.43) and MIVIs (*d*=-.43), and for Arithmetic Reasoning, using SDI (*d*=.43) and MIVIs (*d*=.39), with positive values treated as bias against the focal group (i.e., Black). Adding equality constraints first on latent covariances (M5) and then on the latent variances (M6) produced slightly worse fit at either step. It is possible that the latent variance-covariance matrix is different across groups, which may undermine group comparison to some extent. A more parsimonious model (M7) adds equality constraint on the speed factor mean due to being quite small. It does not deteriorate model fit but the high value of RMSEAD (close to .08) leaves some doubts.

Table 5. Black-White differences among females using Correlated Factors

Model Level	χ2	df	CFI	RMSEA	SRMR	Mc	RMSEAD [CI]
M0. Baseline	70111	499	.950	.043	.032	.629	
M1. Configural	69163	998	.948	.043	.032	.635	
M2. Regression	69589	1032	.948	.042	.033	.633	.016 [.015:.018]
M3. Metric	71432	1073	.946	.042	.034	.626	.032 [.031:.033]
M3a. Partial Metric[1]	70400	1069	.947	.042	.033	.630	.023 [.022:.025]
M4. Scalar	76280	1097	.943	.043	.035	.606	.070 [.069:.072]
M4a. Partial Scalar[2]	72652	1093	.945	.042	.034	.621	.047 [.046:.049]
M5. Lv covariance	74067	1108	.944	.042	.037	.615	.044 [.042:.047]
M6. Lv var-covariance	75159	1114	.943	.042	.038	.610	.062 [.058:.065]
M7. Lv reduced	**75451**	**1115**	**.943**	**.042**	**.038**	**.609**	**.075 [.067:.084]**

[1] Freed parameters (by descending order of χ2 size) are: English =~ Word Functions, Info =~ Health, Math =~ Word Functions, Speed =~ Disguised Words.
[2] Freed parameters (by descending order of χ2 size) are: Mechanics~1, Clerical Checking~1, Arithmetic Computation~1, Arithmetic Reasoning~1.
Note: higher values of CFI and Mc indicate better fit, while lower values of χ2, RMSEA, RMSEAD, SRMR indicate better fit.

Table 6 contains a summary of the fit indices of the HOF model and the free parameters. At the metric level, four loadings showed large χ2, especially the second-order loadings of math and english on *g*. Partial metric allows them to be free, the improvement in fit is minor. At the scalar level, the same four subtests are found to be especially biased. Partial scalar removes their equality constraints and this improves model fit enough. Adding a constraint on the latent variances (M5) does not seem to worsen the model fit. Similarly, the strong SH model fits almost just as well. However, a more parsimonious model (M6b) which adds equality constraints on english, speed, science factor means, fits better than the strong SH, especially when considering the much lower RMSEAD.

Table 6. Black-White differences among females using Higher Order Factor

Model Level	χ2	df	CFI	RMSEA	SRMR	Mc	RMSEAD [CI]
M0. Baseline	79749	508	.944	.046	.037	.590	

M1. Configural	78750	1016	.941	.045	.037	.595	
M2. Regression	79176	1050	.940	.045	.038	.594	.016 [.015:.018]
M3. Metric	82098	1096	.938	.044	.041	.583	.038 [.037:.039]
M3a. Partial Metric[1]	80609	1092	.939	.044	.039	.589	.028 [.027:.029]
M4. Scalar	86643	1119	.935	.045	.040	.565	.071 [.070:.073]
M4a. Partial Scalar[2]	82974	1115	.937	.044	.040	.579	.048 [.047:.050]
M5. Lv variance	83892	1122	.937	.044	.042	.576	.049 [.046:.052]
M6a. Strong SH	84863	1128	.936	.044	.042	.572	.067 [.064:.071]
M6b. Weak SH	**84004**	**1125**	**.937**	**.044**	**.042**	**.575**	**.035 [.030:.040]**

[1] Freed parameters (by descending order of χ2 size) are: English =~ Word Functions, *g* =~ English, *g* =~ Math, Info =~ Health.
[2] Freed parameters (by descending order of χ2 size) are: Mechanics~1, Clerical Checking~1, Arithmetic Computation~1, Arithmetic Reasoning~1.
Note: higher values of CFI and Mc indicate better fit, while lower values of χ2, RMSEA, RMSEAD, SRMR indicate better fit.

Table 7 contains a summary of the fit indices of the BF model and the free parameters. At the metric level, four loadings have by far the largest χ2 values. Partial metric removes their constraints but does not improve model fit: CFI is unchanged and Mc or RMSEAD barely changed. It is possible that the model contains many loadings with small or modest group differences. At this point, freeing more loadings can only undermine group comparison in latent means, despite the assumption of invariant loadings being somewhat ambiguous. At the scalar level, values of ∆CFI=.004 and RMSEAD=.074 are concerning, but this time, the most biased subtests differ a little: mechanics, arithmetic computation, physical science and arithmetic reasoning. Partial scalar releases them, with acceptable improvement in fit. Adding a constraint on the latent variances (M5) does not seem to worsen the model fit. Compared to this model, the strong SH shows worse fit, as shown by Mc and RMSEAD values. The decrement in fit for the no SH is much worse in comparison when judged by RMSEAD. Both models are therefore rejected. A more parsimonious version of model M5 constrained math factor means to be equal, without any change in model fit (M6b).

Table 7. Black-White differences among females using Bifactor

Model Level	χ2	df	CFI	RMSEA	SRMR	Mc	RMSEAD [CI]
M0. Baseline	58518	480	.959	.040	.027	.679	
M1. Configural	57816	960	.957	.040	.028	.684	
M2. Regression	58247	994	.956	.039	.029	.683	.016 [.015:.018]

M3. Metric	61807	1068	.954	.039	.032	.667	.032 [.031:.033]
M3a. Partial Metric[1]	60806	1064	.954	.039	.032	.671	.028 [.027:.029]
M4. Scalar	66601	1091	.950	.040	.033	.646	.074 [.072:.075]
M4a. Partial Scalar[2]	63741	1087	.952	.039	.033	.659	.057 [.055:.058]
M5. Lv variance	64880	1094	.951	.039	.034	.654	.057 [.054:.060]
M6a. Strong SH	66814	1100	.950	.040	.035	.645	.080 [.077:.084]
M6b. Weak SH	**64889**	**1095**	**.951**	**.039**	**.034**	**.654**	**.010 [.003:.020]**
M6c. No SH	68774	1096	.948	.041	.039	.637	.192 [.186:.199]

[1] Freed parameters (by descending order of χ2 size) are: English =~ Word Functions, *g* =~ Word Functions, *g* =~ Health, Speed =~ Disguised Words.
[2] Freed parameters (by descending order of χ2 size) are: Mechanics~1, Arithmetic Computation~1, Physical Science~1, Arithmetic Reasoning~1.
Note: higher values of CFI and Mc indicate better fit, while lower values of χ2, RMSEA, RMSEAD, SRMR indicate better fit.

3.2.3 Robustness analyses

Given the aforementioned issue of power, and following the recommendation of Yoon & Lai (2018), a subsampling approach is used as robustness analysis. The slice_sample() function in R was applied first when constraining all loadings and intercepts to be equal and second when releasing these constraints following the results of the main analysis (shown in Tables 2-7). The random sampling method shows mixed evidence of its efficiency. Upon examining the unstandardized loadings and intercepts in the free model (configural), there are small-modest variations in the loadings but non-trivial variations in the intercepts. Due to this randomness, it is no wonder why some data runs yield worse or better fit for all models (e.g., .006 in ΔCFI). And although there is consistency among the largest biased subtests, upon inspecting the χ2 in modification indices, a few more parameters randomly show up as biased every single run. Their effect on model fit also varies across runs.

Across runs, without releasing any parameters at the metric or scalar level, there is a strong consistency with which the data rank the models. For both the Black-White male and Black-White female groups, the BF model always fits better than the CF model (ΔCFI=.010) which always fits better than the HOF model (ΔCFI=.005). This is a pattern which was apparent in the analysis using the full sample of Whites, but the advantage of the bifactor was too small given the positive bifactor bias reported in recent simulations. Another advantage is the model misfit being more visible. The ΔCFI values for metric and scalar invariance in the bifactor model are .010 and .005 in the male group and .015 and .010 in the female group. The ΔCFI values for metric and scalar invariance in the CF model are .007 and .011 in the male group and .007 and .016 in the female group. Finally, in both of these groups, the superiority of the weak SH (M5 or M6b) over the strong SH or the no SH model is so much clearer within the bifactor, while this pattern was barely visible using the full White sample. One glaring issue comes from measurement invariance, as the effect varies across runs. When attempting to release Mechanics' means (i.e., intercepts) due to being by far the most biased subtest (especially among females), it was found that this single one parameter at partial scalar sometimes causes a change in .001 or .003 in CFI.

Across runs, this time using the constraints applied in the main analyses at each step. This procedure however is much less optimal. Given the randomness of these sliced samples it is expected that the same constraints will not hold even across multiple runs. For three data runs, in the bifactor, partial scalar releases two intercepts but this did not improve model fit at all. Generally, metric and scalar invariance are both strongly rejected even after releasing these constraints, and partial scalar for instance always barely improves model fit in both HOF and BF models. This confirms that the partial constraints applied on the full sample cannot generalize to random sampling. One notable observation here is that the superiority of the bifactor is smaller. This is because the partial scalar usually improves model fit very little in the BF model, as opposed to CF. Modification indices always reveal one (or two) additional subtest(s) of importance in the BF model, but it changes in every single run, and releasing it (or them) substantially improves model fit (ΔCFI=.002-.004).

The above results provide strong evidence that random sampling does not override the conclusion of the main analyses, at least with respect to measurement invariance models. The variability introduced by random sampling should, in principle, amplify group differences in the parameters. If some parameters (regressions, loadings, intercepts, residuals) are equal or almost equal across groups, the added noise in the White group will make invariant parameters non-invariant even if the random noise should average out. This, in turn, affects sensitivity because if all parameters now display some small or modest group differences, then even the most biased parameters will have their impact diluted since now many more parameters are biased to some extent. On the other hand, the model fit indices distinguish between competing models much better and in a consistent way, favoring *g* models (BF vs CF model; Weak SH vs no SH).

The examination of outliers is another, yet necessary, robustness check. All models for both the Black-White male and female groups were rerun after removing multivariate outliers with the Minimum Covariance Determinant (MCD) proposed by Leys et al. (2018) who argued that the basic Mahalanobis Distance was not a robust method. Although the multivariate normality was barely acceptable, the number of outliers was large: MCD removed 1,948 White males and 338 Black males, and 1,005 White females and 372 Black females. The fit indices and parameter estimates in all models barely changed at all (this includes the VPR models as well). If anything has changed, it was the strict invariance model, which somewhat improved in the male group, with CFI=.001, and female group, with CFI=.002. In other words, strict invariance is less violated without outliers.

The final robustness analysis involves a split-half cross-validation which uses random half samples. Although Perry et al. (2015) used a very liberal cutoff of CFI>.01 to determine failed cross-validation, in the current study the difference in CFI value between the original samples and either of the cross-validation samples for any given MGCFA model is at most .001. This indicates that the model fit values are not sample-specific.

3.3 Gender analysis

Overall fit is acceptable in all models, except maybe for Mc. In these analyses, lack of power shouldn't be an issue since there are no serious sample unbalances. Following the criteria suggested by Chen (2007), Khojasteh & Lo (2015), Meade et al. (2008) should therefore be easier than earlier analyses of the Black-White groups. Configural and regression invariance both fit very well. Thus the next levels of invariance will be the focus. Strict invariance is always strongly rejected. Details of these analyses are provided in the supplementary material.

Competing models are evaluated based on their optimal constraints. In the White group, the BF fits marginally better than the CF model, which isn't telling anything due to fit indices slightly favoring the bifactor, whereas the CF model fits largely better than the HOF model. In the Black group, the CF and HOF models fit equally well whereas the BF model fits much better than either of these models. At first glance, this suggests that *g* explains the sex differences among Blacks but not among Whites.

3.3.1 Male-female White group

The model specification is displayed as follows:

```
english =~ S1 + S19 + S20 + S21 + S22 + S23 + S24 + S25 + S26 + S31 + S32
+ S34
math =~ S4 + S5 + S6 + S25 + S32 + S33 + S34
speed =~ S19 + S29 + S34 + S35 + S36 + S37
info =~ S1 + S2 + S3 + S4 + S5 + S7 + S8 + S11 + S12 + S13 + S14 + S15 +
S16 + S19 + S26
science =~ S1 + S6 + S7 + S8 + S9 + S10 + S13 + S28
spatial =~ S28 + S29 + S30 + S31 + S32 + S37
```

Here, it must be noted that two cross-loadings were ignored despite averaging .20 because there would be too many triple loadings otherwise.

Table 8 contains a summary of the fit indices of the CF model and the free parameters. At the metric level, fit deteriorates somewhat but this is still acceptable. Scalar invariance however does not hold. Partial scalar releases the constraints on subtests mean until acceptable fit is achieved. A total of twelve subtests have to be released and yet the RMSEAD of .085 indicates not so good fit, although not critical. Adding constraints on the latent covariances (M5) worsens model fit a little bit but adding constraints on variances (M6) does not change model fit. A more parsimonious model (M7) then adds an equality constraint on information factor means and this fits perfectly.

Table 8. Male-female differences among Whites using Correlated Factors

Model Level	χ2	df	CFI	RMSEA	SRMR	Mc	RMSEAD [CI]
M0. Baseline	156527	492	.945	.047	.039	.578	

M1. Configural	115791	984	.958	.041	.030	.668	
M2. Regression	119145	1018	.957	.040	.031	.660	.035 [.034:.036]
M3. Metric	128322	1066	.953	.041	.039	.639	.048 [.047:.049]
M4. Scalar	195992	1094	.928	.050	.044	.504	.173 [.172:.174]
M4a. Partial Scalar[1]	138034	1082	.950	.042	.039	.618	.085 [.084:.087]
M5. Lv covariance	141168	1097	.948	.042	.055	.611	.051 [.049:.053]
M6. Lv var-covariance	142175	1103	.948	.042	.057	.609	.044 [.041:.046]
M7. Lv reduced	**142176**	**1104**	**.948**	**.042**	**.057**	**.609**	**NaN***

[1] Freed parameters (by descending order of χ2 size) are: Social Studies~1, Theater~1, Law~1, Music~1, Physical Science~1, Miscellaneous~1, Visualization in 3D~1, Health~1, Mechanical Reasoning~1, High School Math~1, Mechanics~1, Art~1.
* NaN is the result of a Chi-square that is negative or lower than 1 (model fits better). RMSEAD therefore cannot be computed.
Note: higher values of CFI and Mc indicate better fit, while lower values of χ2, RMSEA, RMSEAD, SRMR indicate better fit.

Table 9 contains a summary of the fit indices of the HOF model and the free parameters. Similar to the CF model, metric invariance holds while scalar invariance does not. Partial scalar releases the constraint on twelve subtests' means and while ΔCFI=.004 is acceptable, RMSEAD=.089 is borderline. There are two reasons for not freeing more subtests. First, it compromises latent mean comparisons even more. Second, after the release of the most biased subtests (Social Studies, Theater, Law, and Music), each subsequent subtest contributes very little to model improvement, which means reducing RMSEAD to acceptable levels would require many more freed subtests. Next step, latent variance (M5) holds well. Strong SH clearly is rejected. A more parsimonious version of M5 adds an equality constraint on math factor means and this fits perfectly (M6b).

Table 9. Male-female differences among Whites using Higher Order Factor

Model Level	χ2	df	CFI	RMSEA	SRMR	Mc	RMSEAD [CI]
M0. Baseline	211618	501	.926	.054	.056	.476	
M1. Configural	136882	1002	.950	.044	.035	.620	
M2. Regression	140316	1036	.949	.043	.036	.613	.035 [.034:036]
M3. Metric	150122	1089	.945	.044	.043	.592	.047 [.047:.048]
M4. Scalar	215694	1116	.921	.052	.049	.470	.168 [.167:.169]

M4a. Partial Scalar[1]	160436	1104	.941	.045	.044	.571	.089 [.088:.091]
M5. Lv variance	163046	1111	.940	.045	.060	.566	.067 [.065:.069]
M6a. Strong SH	313949	1117	.885	.063	.091	.333	.615 [.613:.618]
M6b. Weak SH	**163046**	**1112**	**.940**	**.045**	**.060**	**.566**	**NaN***

[1] Freed parameters (by descending order of χ2 size) are: Social Studies~1, Theater~1, Law~1, Music~1, Health~1, Miscellaneous~1, Physical Science~1, Visualization in 3D~1, Electronics~1, Math~1, High School Math~1, Biological Science~1.
* NaN is the result of a Chi-square that is negative or lower than 1 (model fits better). RMSEAD therefore cannot be computed.
Note: higher values of CFI and Mc indicate better fit, while lower values of χ2, RMSEA, RMSEAD, SRMR indicate better fit.

Table 10 contains a summary of the fit indices of the BF model and the free parameters. Here, metric invariance holds but upon inspecting modification indices, one loading stands out as having a much larger χ2. Releasing it improves model fit (M3a). Scalar invariance does not hold but releasing 7 subtests allows partial scalar (M4a) to achieve acceptable fit. A constraint on latent variances (M5) perhaps does not hold, as judged by ΔSRMR=.019 and RMSEAD close to .08. A more parsimonious model (M6b) constraining spatial factor means to zero causes worse fit, as judged by RMSEAD. Similarly, both the strong SH (M6a) and no SH (M6c) models are rejected, as judged by all fit indices.

Table 10. Male-female differences among Whites using Bifactor

Model Level	χ2	df	CFI	RMSEA	SRMR	Mc	RMSEAD [CI]
M0. Baseline	168899	473	.941	.050	.052	.553	
M1. Configural	109567	946	.960	.040	.027	.682	
M2. Regression	112939	980	.959	.040	.028	.674	.035 [.034:.036]
M3. Metric	123859	1061	.955	.040	.037	.649	.040 [.039:.040]
M3a. Partial Metric[1]	121921	1060	.956	.040	.036	.654	.036 [.035:.037]
M4. Scalar	151477	1087	.945	.044	.039	.589	.114 [.113:.115]
M4a. Partial Scalar[2]	129982	1080	.953	.041	.037	.635	.066 [.065:.068]
M5. Lv variance	**133080**	**1087**	**.951**	**.041**	**.056**	**.629**	**.074 [.071:.076]**
M6a. Strong SH	303573	1093	.889	.062	.091	.345	.632 [.630:.635]
M6b. Weak SH	133902	1088	.951	.041	.057	.627	.102 [.095:.108]

M6c. No SH	142850	1088	.948	.043	.064	.607	.671 [.665:.677]

[1] Freed parameters (by descending order of χ2 size) are: *g* =~ Vocabulary.
[2] Freed parameters (by descending order of χ2 size) are: Disguised Words~1, Physical Science~1, Health~1, Law~1, Visualization in 3D~1, Aeronautics~1, Biological Science~1.
Note: higher values of CFI and Mc indicate better fit, while lower values of χ2, RMSEA, RMSEAD, SRMR indicate better fit.

Robustness analysis was conducted for the gender analysis in the White group because the multivariate normality was non-normal. Removing outliers (which amounted to 1,918 White males and 1,184 White females) using MCD produced similar parameter estimates and fit indices for all constraints levels and all competing models (including the VPR).

3.3.2 Male-female Black group

The model specification is displayed as follows:

```
english =~ S1 + S7 + S13 + S19 + S20 + S21 + S22 + S23 + S24 + S25 + S26
+ S31 + S34
math =~ S5 + S25 + S32 + S33 + S34
speed =~ S19 + S34 + S35 + S36 + S37
info =~ S1 + S2 + S3 + S4 + S7 + S8 + S10 + S11 + S12 + S13 + S14 + S15 +
S16 + S19 + S26
science =~ S1 + S5 + S6 + S7 + S9 + S10
spatial =~ S28 + S29 + S30 + S31 + S37
```

Table 11 contains a summary of the fit indices of the CF model and the free parameters. As opposed to earlier groups, now metric invariance clearly does not hold (ΔCFI=.008 and RMSEAD=.074). Partial metric releases three loadings, now achieving good fit. Scalar invariance also does not hold. Partial scalar releases seven subtests, reaching acceptable fit despite RMSEAD=.083. Next, the constraints on latent covariances (M5) lead to a serious misfit. Adding then the constraints on latent covariances (M6) leads to minor change CFA and Mc but RMSEAD suggests these constraints may not be acceptable. Overall this means neither the latent covariances or variances seem to be equal across groups. A more parsimonious version of model M6 adds an equality constraint on the information factor means without decreasing model fit.

Table 11. Male-female differences among Blacks using Correlated Factors

Model Level	χ2	df	CFI	RMSEA	SRMR	Mc	RMSEAD [CI]
M0. Baseline	5980	497	.952	.043	.037	.637	
M1. Configural	5342	994	.961	.038	.034	.699	
M2. Regression	5422	1028	.961	.037	.034	.697	.019 [.013:.025]
M3. Metric	6290	1071	.953	.040	.051	.651	.074 [.070:.079]

M3a. Partial Metric[1]	5875	1068	.957	.038	.043	.674	.054 [.049:.059]
M4. Scalar	7468	1096	.943	.044	.053	.592	.130 [.124:.135]
M4a. Partial Scalar[2]	6324	1089	.953	.040	.044	.650	.083 [.077:.090]
M5. Lv covariance	6828	1104	.949	.041	.077	.625	.087 [.079:.095]
M6. Lv var-covariance	6929	1110	.948	.042	.079	.620	.062 [.050:.075]
M7. Lv reduced	**6929**	**1111**	**.948**	**.041**	**.079**	**.620**	**NaN***

[1] Freed parameters (by descending order of χ2 size) are: Info =~ Aeronautics, Spatial =~ Mechanical Reasoning, Info=~ Social Studies.

[2] Freed parameters (by descending order of χ2 size) are: Aeronautics~1, Mechanical Reasoning~1, Mechanics~1, Social Studies~1, Theater~1, Visualization in 3D~1, Physical Science~1.

* NaN is the result of a Chi-square that is negative or lower than 1 (model fits better). RMSEAD therefore cannot be computed.

Note: higher values of CFI and Mc indicate better fit, while lower values of χ2, RMSEA, RMSEAD, SRMR indicate better fit.

Table 12 contains a summary of the fit indices of the HOF model and the free parameters. Here again, metric invariance is strongly violated. Partial metric releases five loadings, producing good fit. It is unclear whether this is sufficient since only ΔCFI=.005 rejects metric invariance, according to Khojasteh & Lo's (2015) cutoffs. Scalar invariance is also rejected. Partial scalar releases six subtests, achieving acceptable fit although RMSEAD is close to .08. Adding a constraint on latent variances (M5) does not worsen model fit, except for SRMR. The Strong SH model is largely rejected. A more parsimonious version of model M5 adds a constraint on math factor means without affecting model fit (M6b).

Table 12. Male-female differences among Blacks using Higher Order Factor

Model Level	χ2	df	CFI	RMSEA	SRMR	Mc	RMSEAD [CI]
M0. Baseline	7120	506	.942	.046	.043	.581	
M1. Configural	5869	1012	.957	.040	.036	.671	
M2. Regression	5951	1046	.956	.039	.037	.668	.020 [.014:.026]
M3. Metric	7194	1094	.945	.043	.064	.606	.085 [.081:.089]
M3a. Partial Metric[1]	6469	1089	.952	.040	.046	.643	.056 [.052:.061]
M4. Scalar	8374	1116	.935	.046	.053	.551	.143 [.138:.149]
M4a. Partial Scalar[2]	6922	1110	.948	.041	.047	.620	.076 [.069:.082]

M5. Lv variance	7009	1117	.947	.042	.065	.616	.053 [.042:.065]
M6a. Strong SH	8675	1123	.933	.047	.076	.537	.305 [.292:.317]
M6b. Weak SH	**7009**	**1118**	**.947**	**.042**	**.065**	**.616**	**NaN***

[1] Freed parameters (by descending order of χ2 size) are: *g* =~ Science, Info =~ Aeronautics, Science =~ Mechanics, *g* =~ English, Spatial =~ Mechanical Reasoning.
[2] Freed parameters (by descending order of χ2 size) are: Mechanics~1, Aeronautics~1, Mechanical Reasoning~1, Physical Science~1, Social Studies~1, Theater~1.
* NaN is the result of a Chi-square that is negative or lower than 1 (model fits better). RMSEAD therefore cannot be computed.
Note: higher values of CFI and Mc indicate better fit, while lower values of χ2, RMSEA, RMSEAD, SRMR indicate better fit.

Table 13 contains a summary of the fit indices of the BF model and the free parameters. Metric invariance is strongly rejected. Partial metric achieves acceptable fit, but only after freeing seven loadings, most of which are the subtests' direct loadings on *g*. Scalar invariance also is rejected. Partial scalar releases four subtests, achieving acceptable fit. However, the constraint on latent variances (M5) is strongly rejected, which is an indication that the groups use different ranges of latent abilities. A Strong SH model (M6a) fits much worse than M5. A more parsimonious version of model M5 which adds an equality constraint on spatial factor means fits a little worse according to CFI only but RMSEAD suggests very good fit (M6b). The no SH model (M6c) fits worse than model M5.

Table 13. Male-female differences among Blacks using Bifactor

Model Level	χ2	df	CFI	RMSEA	SRMR	Mc	RMSEAD [CI]
M0. Baseline	5663	478	.954	.042	.032	.653	
M1. Configural	4623	956	.967	.036	.023	.740	
M2. Regression	4704	990	.967	.035	.024	.737	.020 [.014:.026]
M3. Metric	5911	1066	.957	.039	.059	.671	.063 [.060:.067]
M3a. Partial Metric[1]	5251	1059	.963	.036	.042	.708	.043 [.039:.046]
M4. Scalar	6498	1086	.952	.040	.046	.641	.114 [.108:.119]
M4a. Partial Scalar[2]	5690	1082	.959	.037	.042	.685	.068 [.062:.074]
M5. Lv variance	**6281**	**1089**	**.954**	**.040**	**.064**	**.653**	**.173 [.162:.184]**
M6a. Strong SH	9095	1095	.929	.049	.079	.518	.427 [.415:.439]
M6b. Weak SH	6294	1090	.953	.040	.064	.652	.047 [.021:.080]
M6c. No SH	6554	1091	.951	.041	.072	.638	.248 [.227:.270]

[1] Freed parameters (by descending order of χ2 size) are: *g* =~ Aeronautics, *g* =~ Mechanical Reasoning, *g* =~ Mechanics, Spatial =~ Mechanical Reasoning, *g* =~ Word Functions, *g* =~ Electronics, *g* =~ Social Studies.
[2] Freed parameters (by descending order of χ2 size) are: Table Reading~1, Mechanical Reasoning~1, Social Studies~1, Clerical Checking~1.
Note: higher values of CFI and Mc indicate better fit, while lower values of χ2, RMSEA, RMSEAD, SRMR indicate better fit.

3.4 The contribution of Spearman's *g*

Table 14 shows the group differences in factor means expressed in standardized units,[11] as well as their standard errors, from the best fitting bifactor and best fitting higher order factor models. Bootstrap analysis, using the function *bootstrapLavaan*(), confirmed that the standard errors are accurate. The Black-White *g* gap in the male and female groups are, respectively, 1.5 and 1.3. The male-female *g* gap in the White and Black groups are, respectively, 0.85 and 0.55. The sex gap seems large compared to earlier reports on IQ gaps, until one realizes that this battery of tests has a strong knowledge component, especially specific knowledge, which shows a male advantage (Jensen, 1998, pp. 279, 534, 540). But because the ratio of biased/unbiased subtests was relatively high, the estimates of sex differences in factor means should be interpreted with caution. Moreover, the factor means vary somewhat depending on the specification of loadings and/or cross-loadings.[12] Having the means of the factors is informative but does not tell us how well Spearman's *g* explains the data.

Table 14. *d* gaps (with their S.E.) from the best fitting *g* models per group analysis

	BW *d* (male)		BW *d* (female)		sex *d* (white)		sex *d* (black)	
	BF	HOF	BF	HOF	BF	HOF	BF	HOF
English	–	–	-1.081 (.038)	–	2.816 (.032)	.945 (.015)	1.810 (.089)	.506 (.026)
Math	-.326 (.045)	-.422 (.041)	–	-.237 (.033)	.783 (.021)	–	.808 (.104)	–
Speed	–	–	.225 (.031)	–	.544 (.008)	.426 (.008)	.281 (.048)	.285 (.034)
Information	–	–	-.679	-.290	1.974	.344	1.500	.145

[11] Since the latent means in the reference group must be set at zero for identification, *lavaan* package calculates the fully standardized estimate of the latent means in the focal group by using the focal group's standard deviation. Because the mean of the reference group is zero, the mean of the focal group actually expresses the group standardized difference.
[12] In the bifactor configural model, some loadings were very small (lower than .10). To ensure that the large gender *g* gap is robust to model specification, the invariance models were refit after removing these loadings. This time, the intercepts that had to be freed (based on modification indices) changed slightly compared to the earlier model, which had more cross-loadings. In the best bifactor model, the gender difference in *g* was -.783 and -.449 for the White group and Black group respectively (instead of -.853 and -.554 in Table 14). While the gender gap in *g* changed little, the gender gap in some of the group factors changed drastically. This is because group factor mean differences vary depending on which intercept is fixed.

			(.032)	(.014)	(.024)	(.014)	(.110)	(.025)
Science	-.897 (.032)	-.685 (.024)	-.211 (.033)	–	-1.740 (.016)	-1.895 (.015)	-1.361 (.078)	-.803 (.058)
Spatial	-.430 (.030)	-.374 (.024)	-.783 (.025)	-.516 (.016)	-.329 (.013)	-.559 (.014)	-.179 (.057)	-.466 (.036)
g	-1.502 (.026)	-1.484 (.026)	-1.272 (.020)	-1.315 (.017)	-.853 (.015)	-.324 (.017)	-.554 (.052)	-.150 (.036)

Note: Negative values indicate advantage for Whites (or males).

The proportion of subtest differences due to *g* answers this question more directly. It can be computed, in the case of the bifactor model, by dividing the product of the *g* mean difference and subtest's loading on *g* by the sum of the product of all latent mean differences and their subtest's loadings. This is the method used by Dolan (2000, Table 8 and Eq. 23). However, Dolan multiplied the first-order specific factor means by the first-order loadings (i.e., the path tracing rule) to estimate the *g* loadings due to employing the HOF model. In the BF model, the calculation is easier because there are no first-order factors mediating the relationship between *g* and the subtests.[13] Whatever structure (BF or HOF) is used to compute the proportions, it is important to note that they are not *g*-loadings. To compute the proportions, the loadings of the focal group (Blacks or females) are used.[14]

Table 15 provides the percentage of each subtest mean difference that is due to *g* as opposed to specific factors, based on the best fitting bifactor model by subgroup. For the Black-White groups, many subtests display a very high proportion, close to .8 or 1. The average proportion is .90 for the male group and .73 for the female group, indicating that *g* is the main source of the group differences. At first glance, it may seem puzzling that *g* explains 100% of the Black-White difference in some of the speed subtests in the male group, despite their very low loadings on *g* but high loadings on the speed factor. This is because their mean difference in the speed factor is zero. For the sex groups, on the other hand, the proportions vary greatly in size. Sometimes *g* explains the lion's share of some subtests' mean differences, sometimes *g* explains very little. The average proportion is .43 for the sex group among Whites and .50 for the sex group among Blacks. If SH explicitly states that *g* is the main source of the group difference, it seems that even the weak SH model does not explain well the pattern of sex differences.

Table 15. Proportions of subtest group differences due to *g* based on Bifactor model

Subtests	BW (male)	BW (female)	sex (White)	sex (Black)
S1	0.861	0.780	0.399	0.445
S2	1.000	0.777	0.485	0.470
S3	1.000	0.767	0.402	0.434

[13] Details of the analysis and calculations are provided in the supplementary file.
[14] Some loadings are different across groups, due to non-invariance, among other things, but this does not change the result substantially.

S4	1.000	0.847	0.674	0.550
S5	0.927	1.000	0.699	0.603
S6	0.764	0.801	0.533	0.924
S7	0.789	0.866	0.457	0.556
S8	0.752	0.704	0.296	1.000*
S9	0.613	0.831	0.344	0.419
S10	0.621	0.888	0.304	0.251
S11	1.000	0.702	0.411	0.444
S12	1.000	0.796	0.514	0.554
S13	1.000	0.806	0.465	0.404
S14	1.000	0.861	0.502	0.425
S15	1.000	0.699	0.357	0.382
S16	1.000	0.802	0.574	0.512
S19	1.000	0.539	0.219	0.384
S20	1.000	0.583	0.302	0.304
S21	1.000	0.651	0.326	0.358
S22	1.000	0.689	0.350	0.354
S23	1.000	0.686	0.369	0.349
S24	1.000	0.699	0.373	0.343
S25	0.936	0.832	0.384	0.425
S26	1.000	0.750	0.448	0.447
S28	0.639	0.664	0.405	0.872
S29	0.728	0.548	0.512	0.745
S30	0.756	0.596	0.680	0.730
S31	0.865	0.621	0.418	0.497
S32	0.959	1.000	0.534	0.637
S33	0.842	1.000	0.589	0.536
S34	0.888	0.647	0.362	0.364
S35	1.000	0.675	0.370	0.548

S36	1.000	0.462	0.209	0.239
S37	0.642	0.378	0.211	0.353
Average	0.900	0.734	0.426	0.496

*The real value was actually 1.289, because this subtest's loading on the information factor was negative (-.031) and non-significant (p=.068). If converted to zero instead, the proportion is 1.

Another method used to test SH is MCV. Following te Nijenhuis & van der Flier (1997), correction for unreliability was applied to both the vector of subtests' differences (observed) and *g*-loadings (based on the first unrotated factor). The subtest reliabilities were taken from Major et al. (2012), but these were missing for the speed subtests. For this reason, the analysis is done assuming the reliability of speed subtests is either .60 or .70. The subtests' means and SDs by subgroups are provided in the supplementary file, both in their original metric (without data transformation and normalization) and after normalization and z-score transformed. The *d* gaps are computed based on the original metric of the subtests for this analysis (the result is unchanged when using the z-score transformed data).

The *g*-loadings correlate highly with Black-White *d* gaps but not with sex *d* gaps. After correction for unreliability, the correlations (*g***d*) for the Black-White male, Black-White female, male-female White, and male-female Black groups are, respectively, .79, .79, -.06, -.12. If the reliability for speed subtests is assumed to be .70 instead of .60, the correlations are .80, .80, -.05, -.12. Without applying correction, the correlations are .80, .81, -.03, -.09.

The *d* gaps and *g*-loadings after correction for unreliability are then plotted to detect anomalous patterns using *ggplot2*. For this analysis, the *d* gaps are computed after the subtests have been normalized and z-score transformed. Using the original metric of the variables reveals one anomaly with respect to the Black-White groups: Table Reading is located far below the regression line (underestimated *d*) but not after normalization. This could be because the distribution of this variable is extremely skewed and non-normal (i.e., its median is 12.0 and mean is 13.05 with a non trivial portion of the students scoring around 60-70). None other subtests behave differently depending on the metric that is used.

Upon inspecting the Figures 2-5, the correlation of *g* with Black-White or sex difference is somewhat affected by the speeded subtests. Removing them yield a correlation *g***d* of .465, .623, -.320, -.427 for Black-White male, Black-White female, male-female White, male-female Black, respectively, because their removal results in a much narrower distribution of *g*-loadings, which negatively affects correlations. Furthermore, some subtests show a larger (or smaller) *d* gap than what is expected based on their *g*-loadings. Mechanics and Mechanical Reasoning are placed well above the regression line, indicating overestimation of *d*, and Clerical Checking well below the regression line, indicating underestimation of *d*, in all subgroup differences. These subtests are obviously poorly explained by *g*. This does not mean they are necessarily biased, as MCV is not designed to detect bias. The unexplained factors could be due to subtests' uniqueness or cultural bias, unless psychometric bias was already accounted for by removing the offending subtests prior to MCV test, leaving uniquenesses (and perhaps measurement error) as the only non-*g* sources. Another reason to be cautious in interpreting the source of *d* is that the correlation, and therefore the direction of the regression line, depends on the inclusion of the speeded subtests.

Figure 2. Regression of standardized difference on *g* loadings among Black-White males

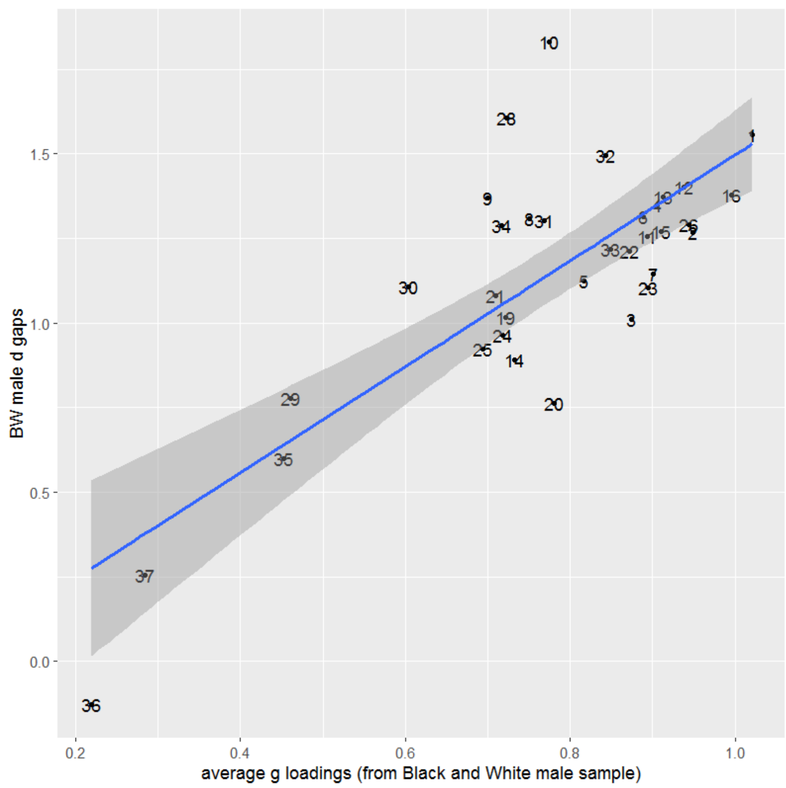


Figure 3. Regression of standardized difference on *g* loadings among Black-White females

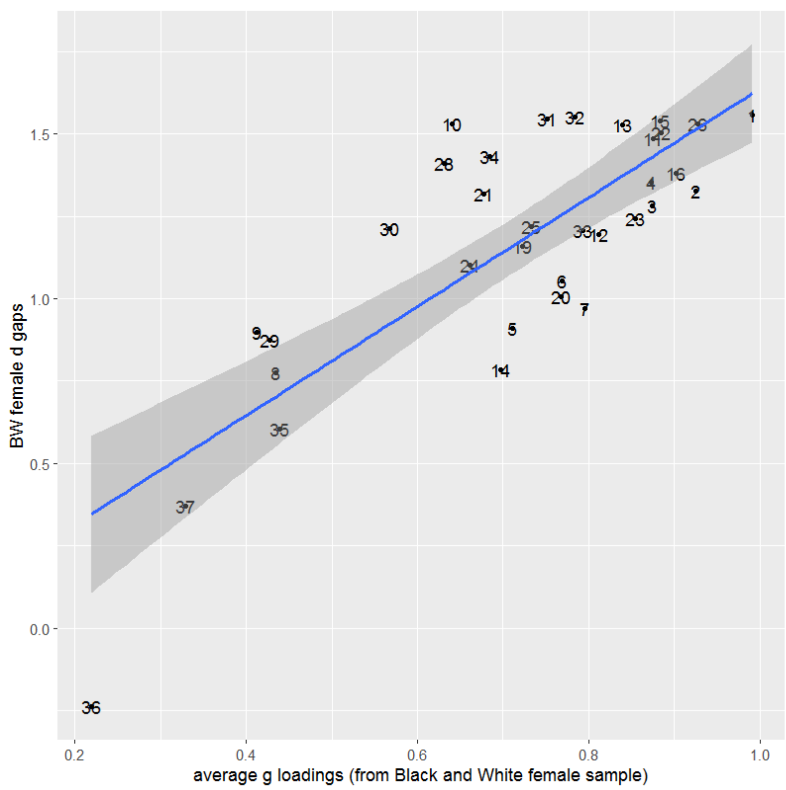


Figure 4. Regression of standardized difference on *g* loadings among male-female Whites

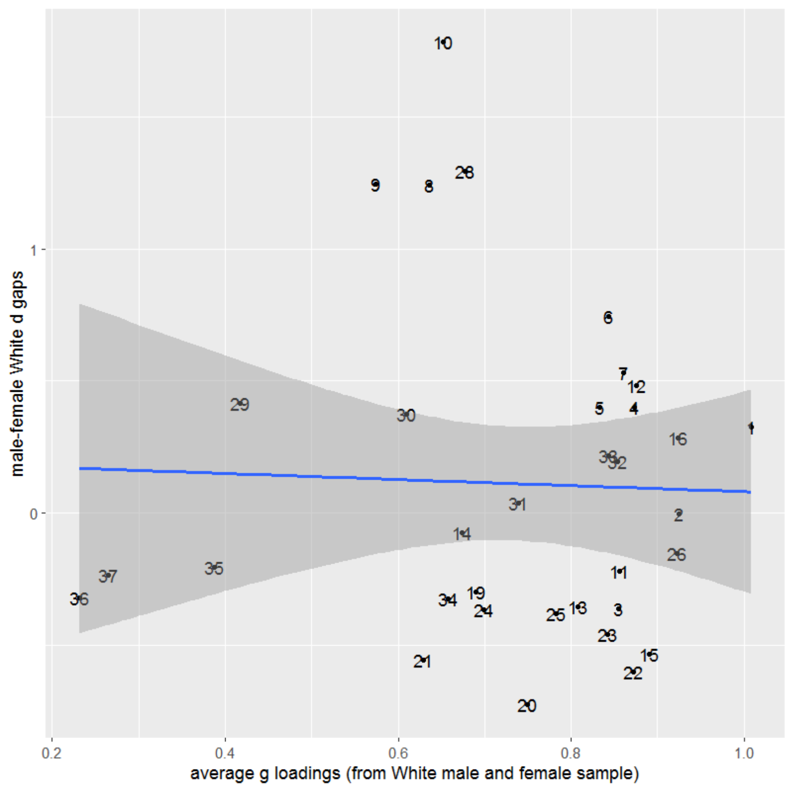
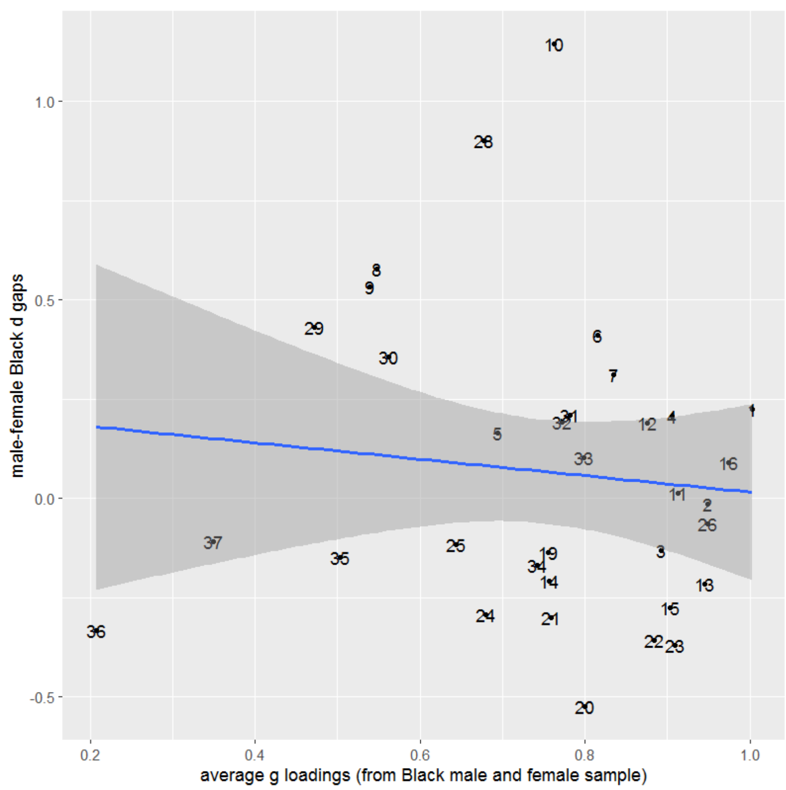


Figure 5. Regression of standardized difference on *g* loadings among male-female Blacks



Gender differences in the subtests' means are smaller than racial differences, with one exception. Some science subtests, along with Mechanical Reasoning, show very large *d* gaps, much larger than expected based on their *g*-loadings. Interestingly, MGCFA revealed these subtests' means as biased with respect to gender in either the White or the Black group.

It is possible that the presence of bias could have affected the correlations. MGCFA analysis revealed the subtests Mechanics, Physical Science, Arithmetic Reasoning and Arithmetic Computation to be biased at the intercept level with respect to racial groups; these were removed prior to MCV test. The correlations improved for the Black-White male and Black-White female groups, going up to .845 and .837, respectively. So far this is consistent with te Nijenhuis et al.'s (2016) conclusion that subtest bias could negatively affect the correlations. The subtest Clerical Checking was also biased (against Whites) but removing it thereafter drops the correlations to .779 and .787.

A similar procedure was done for testing gender differences. After removing all subtests with intercept bias, the correlation is barely affected in both the White and Black groups (regardless of whether one removes the subtest biases found in the HOF or BF model). The exception is for the Black group when removing subtest biases based on the BF model; the negative correlation amplifies (r=-.304) but this is simply because of the removal of two speed subtests. In other words, there is no evidence that biases affect the correlations for the gender differences.

A notable detail in the estimated gender difference in *g* between MCV and MGCFA is that MCV predicts a gender gap close to zero, regardless of the *g*-loading range. This is inconsistent with results provided by MGCFA but not necessarily surprising because these methods address different questions. For instance, MCV predicts a near-zero gender subtest gap for a subtest with a *g*-loading close to 1, whereas MGCFA estimates a gender gap of approximately 0.55 among Blacks and 0.85 among Whites on the latent *g* factor. On the other hand, MCV also predicts a Black-White subtest gap for a *g*-loading close to 1 that closely aligns with the latent *g* factor score gaps estimated by MGCFA.

4. Discussion

The present analysis replicates a pattern that is often observed in previous analyses using MGCFA: that the Black-White cognitive gap is relatively unbiased whereas the sex cognitive gap sometimes exhibits a high percentage of biased subtests. In the Project Talent, the subtests' biases often favoring White students whereas the biases seem to cancel out between sex groups in all models,[15] the exception being the bifactor in which all the biased subtests favor White female students. The number of biased subtests' intercepts is small in the Black-White sample (2 among male and 4 among female groups). The effect size in non-invariance indicates a non-trivial bias, mostly favoring Whites, but given the small ratio of biased/unbiased tests, the effect on the total score should not be large. It is not impossible however, owing to lack of power due to large model size and sample unbalances, that a few

---

[15] Effect sizes are not computed for the sex groups. The conclusion that biases may cancel out is based on the observation of their unstandardized intercepts. Their real magnitude is unknown.

more parameters need to be released in the Black-White female group analysis (either loadings or intercepts or both). If this is the case, the racial bias cannot be considered as minor anymore. It is however unclear whether traditional MGCFA actually lacks sensitivity overall or is too strict in its assumption. Some researchers recently recognized that the assumption of exact equality rather than approximate equality makes MGCFA too strict for many applications, which is the very reason why most studies fail to achieve scalar or strict invariance in survey scales. Van De Schoot et al. (2015), in discussing the strength of approximate measurement invariance (MI) methods such as the Bayesian SEM, wrote: "If there are many small differences between the groups in terms of intercepts or factor loadings, approximate MI seeks a balance between adherence to the requirements of MI, making comparisons possible, and obtaining a well-fitting model (i.e., a model that is more realistic given the data at hand)."

On the other hand, one may argue that releasing more parameters to assess their effect sizes can reveal potential biases, especially when power is suspect (Lasker et al., 2021). There are merits but also difficulties with this approach, since the effect sizes available have limitations in their applications. The proposed effect sizes (SDI or MIVIs) rest on several assumptions: that the intercept's effect size being computed must also assume equal loading and that the model does not contain any cross loadings. According to Nye & Drasgow (2011), the standard formula to compute effect size does not account for cross-loadings. Another complication outlined by Groskurth (2023) and Millsap & Olivera-Aguilar (2012) is that the calculation of non-invariant intercepts typically assumes invariant loadings. It isn't to say there is no way to calculate effect sizes under these conditions, but that the interpretation is less ambiguous in the absence of cross loadings and/or non-invariance in loadings. Because these effect sizes were designed initially for very simple models, the effect sizes of parameter bias reported in this study should be taken with caution.

After establishing partial invariance, SH was tested in all subgroups. This was validated in the Black-White analyses based on two findings: 1) non-*g* models fit worse than *g*-models and 2) the proportion of the subtests' mean differences due to *g* is very large. This was not found to be the case in the gender analyses, although the number of freed parameters undermines group comparison in latent means a little bit. The pattern of sex differences in latent means is worth discussing. At first glance, the large sex difference in *g* scores looks suspicious, given past studies using MGCFA. One must keep in mind that this battery of tests requires a great deal of knowledge, especially specific knowledge for some subtests. This means that *g* in this battery is contaminated by a strong knowledge component, as opposed to standard IQ tests.

MCV was applied to check the similarity of obtained results with MGCFA. The finding of a large correlation between the Black-White *d* gaps and *g*-loadings is consistent with earlier reports on Black adults (te Nijenhuis & Van den Hoek, 2016). It is worth noting that MCV was not meant to assess test bias and is not a latent variable approach. Differences with MGCFA are expected (Lasker et al., 2021). Taken individually, MGCFA is a more complete and reliable method for testing SH as well as test bias, but it has been argued that the consistency of MCV can be improved using meta-analytic correction for artifacts (te Nijenhuis et al., 2019). Unlike MGCFA, the MCV has not been widely accepted due to continuous criticism. For a current state of the debate, see te Nijenhuis et al. (2019).

While MGCFA reveals a clear superiority of the bifactor over the correlated factors model, when using the random sampling approach, the higher order factor usually fits worse than the correlated factors model. This ambiguous finding is important because some researchers have suggested that if parsimony is a desirable outcome, the higher order factor model should be preferred over the bifactor due to having more degrees of freedom. But the debate is not settled. Conceptually the bifactor can be thought as more parsimonious than the higher order factor model at explaining the relationship between subtests and *g* because it does not require a theoretical justification for full mediation of the specific factors and does not impose proportionality constraints on the loadings, despite the bifactor model having fewer degrees of freedom due to introducing more parameters (Cucina & Byle, 2017; Gignac, 2006a, 2006b, 2008). Perhaps more importantly, a bifactor model is consistent with Spearman's initial conceptualization of *g* as having direct influences on the measured tests (Frisby & Beaujean, 2015, p. 95).

When it comes to modeling the Spearman's Hypothesis, one must bear in mind that even if non-*g* models fit the data equally well as *g* models, the former can hardly explain the correlations between *g*-loadings and cognitive differences or the ubiquitous role of *g* at explaining brain mechanisms and cognitive processes (Jensen, 1998). Regardless, it is no less important to acknowledge the totality of the evidence by considering alternative tests of the Spearman's Hypothesis (Jensen, 1985). For instance, by 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 than the FDS gap (Ganzach, 2016; Jensen, 1998, p. 370). Perhaps the most powerful and direct way of testing SH is by examining ECT's task complexity. Jensen (1998, p. 391) reported high correlation between task complexity and the magnitude of the Black-White gap (r=.86). Although SH has been widely replicated using various strategies, misunderstanding or misportrayal of the theory's assumption leads to flawed study designs and ambiguous conclusions. A few examples are provided below.

An argument can be made that a proper test of SH is not possible if cultural and cognitive complexity covary (Helms-Lorenz et al., 2003; Malda et al., 2010). However it is cultural bias and not cultural load which undermines group comparison. A culture loaded item or test is biased only if the groups are differentially exposed to the specific knowledge elicited by the test, given equal latent ability. Without establishing causality first, the argument of cultural confound is not even valid to begin with. Yet this cultural confound is the reason put forth by Helms-Lorenz et al. (2003) as for why they did not find a relationship between *g*-loadings and group differences contrasting second-generation migrants with the majority group. te Nijenhuis & van der Flier (2003) showed that they used a convenience sample, as reflected in the extreme variations in their reported effect sizes, and did not evaluate test bias prior to testing SH. By employing a representative sample, and after removing the strong bias introduced by the Vocabulary subtest, using an extrapolated regression line technique, te Nijenhuis & van der Flier (2003) found a positive correlation between *g*-loadings and group differences. Another issue with the Helms-Lorenz et al.'s (2003) study, which hasn't been pointed out yet, is that cultural loading was rated by psychology students but the criteria were not even defined. Jensen (1980, pp. 570, 637, 639-640) stated that the magnitude of cultural loading is best defined by the rarity of words or rarity of informational content.

Not properly defining cultural loading or cognitive loading, or both, leads to serious flaws in study designs as well. Malda et al.'s (2010) study serves as a striking example. Their analysis may not have removed the influence of cultural bias on cultural load since item bias detection was assessed with a sub-optimal method to detect Differential Item Functioning (DeMars, 2010), a logistic regression effect size measure which is known to underestimate DIF (Oliveri et al., 2012). As a result, they found that cultural complexity rather than cognitive complexity explains the cognitive gap between the Black (urban and rural Tswana) and White (urban Afrikaans) South African groups. But their test of Spearman's against the cultural hypothesis was highly suspect to begin with. Instead of measuring cultural complexity by the rarity of words or content, they measure it by the group difference in familiarity. Because it was defined as differential exposure, cultural complexity here is an index of culture bias, not culture load. Instead of measuring cognitive complexity by the items' complexity within the test using a latent variable approach, they measure it by arbitrarily ranking the complexity between tests, with memory and attention tests assumed to be low in complexity and reasoning tests to be high in complexity. Given such odd procedures, any inference about SH is at best ambiguous.

What if these studies were actually correct and cultural load and cognitive complexity were correlated? A popular idea is that cultural load is necessarily undesirable and must be reduced to zero. As te Nijenhuis & van der Flier (2003) expressed clearly, cultural loading is unavoidable and even desirable as long as future school and work achievement may have a high cultural loading. Removing such items and/or subtests may adversely affect the predictive validity of the test.

A common misconception about verbal tests, often the target of criticism, is that they must always contain a high degree of cultural content. In criticizing Jensen & McGurk (1987) study, which found a stronger Black-White gap on nonverbal tests than verbal tests, Helms-Lorenz et al. (2003) concluded that "An inspection of the items that were rated as least cultural, such as verbal analogies, verbal opposites, and clock problems, suggests that at least some of the items contain fairly strong cultural elements." (p. 11). But as Jensen (1980) noted a long time ago, "verbal analogies based on highly familiar words, but demanding a high level of relation eduction are loaded on gf, whereas analogies based on abstruse or specialized words and terms rarely encountered outside the context of formal education are loaded on gc." (p. 234).

Even if we accept the idea that cultural and cognitive complexity are correlated, there are multiple reasons for rejecting the hypothesis of culture loading as the source of *g* differences. First, Jensen (1998, p. 89) argued that culturally loaded items such as vocabulary require a great deal of fluid ability because most words in a person's vocabulary are learned through inferences of their meaning by the eduction of relations and correlates. The higher the level of a person's *g*, the fewer encounters with a word are needed to correctly infer its meaning. Knowledge gap is a necessary outcome of a *g* gap rather than the opposite (Jensen, 1973, pp. 89-90; 1980, pp. 110, 235). This proposition is fully supported by Jensen's (1973) observation that "An interesting difference between scholastic achievement scores and intelligence test scores (including vocabulary) is that the latter go on increasing steadily throughout the summer months while the children are not in school, while there is an actual loss in achievement test scores from the beginning to the end of the summer." (pp. 90-91). A finding later confirmed by a meta-analytic review (Cooper et al., 1996). Second, the analysis comparing normal-hearing with deaf people, which serves as a quasi-experimental study of the cultural effects on IQ, reveals that only verbal IQ but not performance IQ on the Wechsler was severely deprived as a result of social isolation and non supportive interactions (Braden, 1994; Hu, 2014). That performance IQ is perfectly intact is an indication that cultural deprivation affects domain-specific rather than general ability. Cultural advantage does not result in a higher *g*.

The great majority of the studies confirms the comparability of IQ tests between groups of different demographics and cultures, the exception mainly comes from South African samples (Dolan et al., 2004; Lasker, 2021). Due to the omnipresent force of the mass-market culture in developed countries, it is not surprising that culture bias is rarely noticeable (Rowe et al. 1994, 1995). What is surprising is the conclusion that IQ tests may be more biased with respect to gender than racial groups.

Attempts to reduce the racial IQ gap using alternative cognitive tests have always been proposed (Jensen, 1973, pp. 299-301; 1980, pp. 518, 522). The most recent, but unconvincing, attempt at reducing the cognitive gap comes from Goldstein et al. (2023). They devised a reasoning test composed of novel tasks that do not require previously learned language and quantitative skills. Because they found a Black-White *d* gap ranging between 0.35 and 0.48 across their 6 independent samples, far below the typically found *d* gap of 1.00, they concluded that traditional IQ tests are biased. First, they carefully ignore measurement invariance studies. Second, traditional IQ tests were not administered alongside to serve as benchmarks. Third, their analysis adjusted for socio-economic status because they compare Blacks and Whites who had the same jobs (police officers, deputy sheriffs, firefighters) within the same cities. This study reflects the traditional view that IQ tests are invalid as long as they contain even the slightest cultural component.

The Project Talent administered aptitude tests. They serve as a proxy for cognitive tests, but they are not cognitive tests. For instance, most of the information test items require specific knowledge: asking who was the hero of the Odyssey, or what a female horse is called, etc. They do not call for relation eduction. Jensen (1985) himself has been highly critical of the Project Talent test battery: “Many of these tests are very short, relatively unreliable, and designed to assess such narrow and highly culture-loaded content as knowledge about domestic science, farming, fishing, hunting, and mechanics” (p. 218). Other tests, fortunately, require inductive reasoning and the use of knowledge to find solutions to new problems, in a way consistent with Jensen’s (1973, p. 75) idea of intelligence as being reflected by the broad transfer of the learning in new relevant situations. Overall, this is a mixed bag. But what the present study shows, together with previous studies on test bias in other aptitude tests (Drasgow et al., 2010;[16] Hu et al., 2019; Lasker et al., 2021), is that aptitude tests produce similar outcomes to cognitive tests and that measurement equivalence as well as Spearman’s *g* have been repeatedly confirmed for racial differences but not gender differences.

References

[16] These authors provided evidence of measurement equivalence in the AFOQT aptitude battery across gender and race, but all 4 racial categories were analyzed simultaneously and CFI was reported with 2 decimals.

Arribas-Aguila, D., Abad, F. J., & Colom, R. (2019). Testing the developmental theory of sex differences in intelligence using latent modeling: Evidence from the TEA Ability Battery (BAT-7). *Personality and Individual Differences, 138*, 212–218. doi: 10.1016/j.paid.2018.09.043
Beaujean, A. A., & McGlaughlin, S. M. (2014). Invariance in the Reynolds Intellectual Assessment Scales for Black and White referred students. *Psychological Assessment, 26*(4), 1394–1399. doi: 10.1037/pas0000029
Beaujean, A. A., Parkin, J., & Parker, S. (2014). Comparing Cattell–Horn–Carroll factor models: Differences between bifactor and higher order factor models in predicting language achievement. *Psychological Assessment, 26*(3), 789–805. doi: 10.1037/a0036745
Benson, N., Kranzler, J. H., & Floyd, R. G. (2020). Exploratory and Confirmatory Factor Analysis of the Universal Nonverbal Intelligence Test–Second Edition: Testing Dimensionality and Invariance Across Age, Gender, Race, and Ethnicity. *Assessment, 27*(5), 996–1006. doi: 10.1177/1073191118786584
Bornovalova, M. A., Choate, A. M., Fatimah, H., Petersen, K. J., & Wiernik, B. M. (2020). Appropriate use of bifactor analysis in psychopathology research: Appreciating benefits and limitations. *Biological psychiatry, 88*(1), 18–27. doi: 10.1016/j.biopsych.2020.01.013
Braden, J. P. (1994). *Deafness, deprivation, and IQ*. Springer Science & Business Media. doi: 10.1007/978-1-4757-4917-5
Cao, C., & Liang, X. (2023). The Impact of Ignoring Cross-loadings on the Sensitivity of Fit Measures in Measurement Invariance Testing. *Structural Equation Modeling: A Multidisciplinary Journal*, 1–17. 10.1080/10705511.2023.2223360
Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(3), 464–504. doi: 10.1080/10705510701301834
Chen, H., Zhang, O., Raiford, S. E., Zhu, J., & Weiss, L. G. (2015). Factor invariance between genders on the Wechsler intelligence scale for children–fifth edition. *Personality and Individual Differences, 86*, 1–5. doi: 10.1016/j.paid.2015.05.020
Chen, H., Zhu, J., Liao, Y. K., & Keith, T. Z. (2020). Age and gender invariance in the Taiwan Wechsler intelligence scale for children: Higher order five-factor model. *Journal of Psychoeducational Assessment, 38*(8), 1033–1045. doi: 10.1177/0734282920930542
Cheung, G. W., & Rensvold, R. B. (2001). The effects of model parsimony and sampling error on the fit of structural equation models. *Organizational Research Methods, 4*(3), 236–264. doi: 10.1177/109442810143004
Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*(2), 233–255. doi: 10.1207/s15328007sem0902_5
Cockcroft, K., Alloway, T., Copello, E., & Milligan, R. (2015). A cross-cultural comparison between South African and British students on the wechsler adult intelligence scales third edition (WAIS-III). *Frontiers in Psychology, 6*, 297. doi: 10.3389/fpsyg.2015.00297
Cooper, H., Nye, B., Charlton, K., Lindsay, J., & Greathouse, S. (1996). The Effects of Summer Vacation on Achievement Test Scores: A Narrative and Meta-Analytic Review. *Review of Educational Research, 66*(3), 227–268. doi: 10.3102/00346543066003227
Cucina, J., & Byle, K. (2017). The bifactor model fits better than the higher-order model in more than 90% of comparisons for mental abilities test batteries. *Journal of Intelligence, 5*(3), 27. doi: 10.3390/jintelligence5030027

Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods, 1*(1), 16–29. doi: 10.1037/1082-989x.1.1.16
DeMars, C. E. (2010). Type I error inflation for detecting DIF in the presence of impact. *Educational and Psychological Measurement, 70*(6), 961-972. doi: 10.1177/0013164410366691
DeShon, R. P. (2004). Measures are not invariant across groups without error variance homogeneity. *Psychology Science, 46*, 137-149.
Dolan, C. V. (2000). Investigating Spearman's hypothesis by means of multi-group confirmatory factor analysis. *Multivariate Behavioral Research, 35*(1), 21–50. doi: 10.1207/S15327906MBR3501_2
Dolan, C. V., Colom, R., Abad, F. J., Wicherts, J. M., Hessen, D. J., & Van De Sluis, S. (2006). Multi-group covariance and mean structure modeling of the relationship between the WAIS-III common factors and sex and educational attainment in Spain. *Intelligence, 34*(2), 193–210. doi: 10.1016/j.intell.2005.09.003
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. doi: 10.1016/j.intell.2003.09.001
Dombrowski, S. C., Watkins, M. W., McGill, R. J., Canivez, G. L., Holingue, C., Pritchard, A. E., & Jacobson, L. A. (2021). Measurement Invariance of the Wechsler Intelligence Scale for Children, 10-Subtest Primary Battery: Can Index Scores be Compared across Age, Sex, and Diagnostic Groups?. *Journal of Psychoeducational Assessment, 39*(1), 89–99. doi: 10.1177/0734282920954583
Drasgow, F., Nye, C. D., Carretta, T. R., & Ree, M. J. (2010). Factor structure of the Air Force Officer Qualifying Test Form S: Analysis and comparison with previous forms. *Military Psychology, 22*(1), 68–85. doi: 10.1080/08995600903249255
Fan, X., & Sivo, S. A. (2009). Using Δgoodness-of-fit indexes in assessing mean structure invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 16*(1), 54–69. doi: 10.1080/10705510802561311
Flanagan, J. C., Dailey, J. T., Shaycoft, M. F., Gorham, W. A., Orr, D. B., & Goldberg, I. (1962). *Design for a study of American youth*. Boston: Houghton Mifflin.
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. doi: 10.1016/j.intell.2015.04.007
Ganzach, Y. (2016). Another look at the Spearman's hypothesis and relationship between Digit Span and General Mental Ability. *Learning and Individual Differences, 45*, 128–132. doi: 10.1016/j.lindif.2015.12.026
Gignac, G. E. (2005). Revisiting the factor structure of the WAIS-R: Insights through nested factor modeling. *Assessment, 12*(3), 320–329. doi: 10.1177/1073191105278118
Gignac, G. E. (2006a). A confirmatory examination of the factor structure of the Multidimensional Aptitude Battery: Contrasting oblique, higher order, and nested factor models. *Educational and Psychological Measurement, 66*(1), 136–145.
Gignac, G. E. (2006b). The WAIS-III as a nested factors model: A useful alternative to the more conventional oblique and higher-order models. *Journal of Individual Differences, 27*(2), 73–86. doi: 10.1027/1614-0001.27.2.73

Gignac, G. E. (2008). Higher-order models versus direct hierarchical models: g as superordinate or breadth factor?. *Psychology Science, 50*(1), 21–43.
Goldstein, H. W., Yusko, K. P., Scherbaum, C. A., & Larson, E. C. (2023). Reducing Black–White Racial Differences on Intelligence Tests Used in Hiring for Public Safety Jobs. *Journal of Intelligence, 11*(4), 62. doi: 10.3390/jintelligence11040062
Greene, A. L., Eaton, N. R., Li, K., Forbes, M. K., Krueger, R. F., Markon, K. E., ... & Kotov, R. (2019). Are fit indices used to test psychopathology structure biased? A simulation study. *Journal of Abnormal Psychology, 128*(7), 740–764. doi: 10.1037/abn0000434.supp
Groskurth, K. (2023). Why one size does not fit all: Evaluating the validity of fixed cutoffs for model fit indices and developing new alternatives. *Doctoral dissertation*.
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. doi: 10.1016/j.jsp.2022.05.001
Helms-Lorenz, M., Van de Vijver, F. J., & Poortinga, Y. H. (2003). Cross-cultural differences in cognitive performance and Spearman's hypothesis: g or c?. *Intelligence, 31*(1), 9–29. doi: 10.1016/s0160-2896(02)00111-3
Hsu, H. Y., Skidmore, S. T., Li, Y., & Thompson, B. (2014). Forced zero cross-loading misspecifications in measurement component of structural equation models. *Methodology*. doi: 10.1027/1614-2241/a000084
Hu, M. (2014). The study of deaf people since Braden (1994). *Human Varieties*. Retrieved from: https://humanvarieties.org/2014/09/21/the-study-of-deaf-people-since-braden-1994/
Hu, M. (2023). On The Validity of The GSS Vocabulary Test Across Groups. *OpenPsych*. doi: 10.26775/op.2023.06.22
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. doi: 10.3390/psych1010017
Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1–55. doi: 10.1080/10705519909540118
Irwing, P. (2012). Sex differences in g: An analysis of the US standardization sample of the WAIS-III. *Personality and individual differences, 53*(2), 126–131.doi: 10.1037/e676392012-034
Jensen, A. R. (1973). *Educability and group differences*. New York: Harper & Row.
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. doi: 10.1017/s0140525x00020392
Jensen, A. R. (1998). *The g factor: The science of mental ability*. Westport, CT: Prager.
Jensen, A. R., & McGurk, F. C. J. (1987). Black-white bias in 'cultural' and 'noncultural' test items. *Personality and Individual Differences, 8*(3), 295–301. doi: 10.1016/0191-8869(87)90029-8
Johnson, W., & Bouchard, T. J. (2007). Sex differences in mental ability: A proposed means to link them to brain structure and function. *Intelligence, 35*(3), 197–209. doi: 10.1016/j.intell.2006.07.003
Kane, H. D., & Oakland, T. D. (2010). Group Differences in Cognitive Ability: A CHC Theory Framework. *Mankind Quarterly, 50*(4), 318–331. doi: 10.46469/mq.2010.50.4.4
Keith, T. Z., Fugate, M. H., DeGraff, M., Diamond, C. M., Shadrach, E. A., & Stevens, M. L. (1995). Using Multi-Sample Confirmatory Factor Analysis to Test for Construct Bias: An

Example Using the K-ABC. *Journal of Psychoeducational Assessment, 13*(4), 347–364. doi: 10.1177/073428299501300402
Keith, T. Z., & Reynolds, M. R. (2018). Using confirmatory factor analysis to aid in understanding the constructs measured by intelligence tests. In D. P. Flanagan & E. M. McDonough (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (4th ed., pp. 853–900). The Guilford Press.
Keith, T. Z., Reynolds, M. R., Roberts, L. G., Winter, A. L., & Austin, C. A. (2011). Sex differences in latent cognitive abilities ages 5 to 17: Evidence from the Differential Ability Scales—Second Edition. *Intelligence, 39*(5), 389–404. doi: 10.1016/j.intell.2011.06.008
Khojasteh, J., & Lo, W. J. (2015). Investigating the sensitivity of goodness-of-fit indices to detect measurement invariance in a bifactor model. *Structural Equation Modeling: A Multidisciplinary Journal, 22*(4), 531–541. doi: 10.1080/10705511.2014.937791
Lasker, J. (2021). Interpreting Cross-Cultural Bias in Psychological Assessments: An Empirical Example. doi: 10.31234/osf.io/zwb4c
Lasker, J., Nyborg, H., & Kirkegaard, E. O. W. (2021). Spearman's Hypothesis in the Vietnam Experience Study and National Longitudinal Survey of Youth '79. doi: 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. doi: 10.3390/psych1010034
Lemos, G. C., Abad, F. J., Almeida, L. S., & Colom, R. (2013). Sex differences on g and non-g intellectual performance reveal potential sources of STEM discrepancies. *Intelligence, 41*(1), 11–18. doi: 10.1016/j.intell.2012.10.009
Leys, C., Klein, O., Dominicy, Y., & Ley, C. (2018). Detecting multivariate outliers: Use a robust variant of the Mahalanobis distance. *Journal of Experimental Social Psychology, 74*, 150–156.
Little, T. D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioral Research, 32*(1), 53–76. doi: 10.1207/s15327906mbr3201_3
Little, T. D. (2013). *Longitudinal structural equation modeling*. Guilford press.
Lubke, G. H., & Dolan, C. V. (2003). Can unequal residual variances across groups mask differences in residual means in the common factor model?. *Structural Equation Modeling, 10*(2), 175–192. doi: 10.1207/s15328007sem1002_1
Lubke, G. H., Dolan, C. V., Kelderman, H., & Mellenbergh, G. J. (2003). On the relationship between sources of within-and between-group differences and measurement invariance in the common factor model. *Intelligence, 31*(6), 543–566. doi: 10.1016/s0160-2896(03)00051-5
Major, J. T., Johnson, W., & Deary, I. J. (2012). Comparing models of intelligence in Project TALENT: The VPR model fits better than the CHC and extended Gf–Gc models. *Intelligence, 40*(6), 543–559. doi: 10.1016/j.intell.2012.07.006
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. doi: 10.1016/j.intell.2010.07.004
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. doi: 10.1111/ijsa.12081
Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology, 93*(3), 568–592. doi: 10.1037/0021-9010.93.3.568

Millsap, R. E., & Olivera-Aguilar, M. (2012). Investigating measurement invariance using confirmatory factor analysis In: Hoyle, R. H. (Ed.), *Handbook of Structural Equation Modeling* (pp. 380-292). Guilford Press.
Morgan, G. B., Hodge, K. J., Wells, K. E., & Watkins, M. W. (2015). Are fit indices biased in favor of bi-factor models in cognitive ability research?: A comparison of fit in correlated factors, higher-order, and bi-factor models via Monte Carlo simulations. *Journal of Intelligence, 3*(1), 2–20. doi: 10.3390/jintelligence3010002
Murray, C. (2021). *Facing reality: Two truths about race in America*. Encounter Books.
Murray, A. L., & Johnson, W. (2013). The limitations of model fit in comparing the bi-factor versus higher-order models of human cognitive ability structure. *Intelligence, 41*(5), 407–422. doi: 10.1016/j.intell.2013.06.004
Muthén, B., & Asparouhov, T. (2014). IRT studies of many groups: The alignment method. *Frontiers in Psychology, 5*, 978. doi: 10.3389/fpsyg.2014.00978
Nye, C. D., & Drasgow, F. (2011). Effect size indices for analyses of measurement equivalence: Understanding the practical importance of differences between groups. *Journal of Applied Psychology, 96*(5), 966–980. doi: 10.1037/a0022955
Oliveri, M. E., Olson, B. F., Ercikan, K., & Zumbo, B. D. (2012). Methodologies for Investigating Item- and Test-Level Measurement Equivalence in International Large-Scale Assessments. *International Journal of Testing, 12*(3), 203–223. doi: 10.1080/15305058.2011.617475
Palejwala, M. H., & Fine, J. G. (2015). Gender differences in latent cognitive abilities in children aged 2 to 7. *Intelligence, 48*, 96-108. doi: 10.1016/j.intell.2014.11.004
Pauls, F., Daseking, M., & Petermann, F. (2020). Measurement Invariance Across Gender on the Second-Order Five-Factor Model of the German Wechsler Intelligence Scale for Children–Fifth Edition. *Assessment, 27*(8), 1836–1852. doi: 10.1177/1073191119847762
Perry, J. L., Nicholls, A. R., Clough, P. J., & Crust, L. (2015). Assessing Model Fit: Caveats and Recommendations for Confirmatory Factor Analysis and Exploratory Structural Equation Modeling. *Measurement in Physical Education and Exercise Science, 19*(1), 12–21. doi: 10.1080/1091367X.2014.952370
Pezzuti, L., & Orsini, A. (2016). Are there sex differences in the Wechsler Intelligence Scale for Children — Forth Edition? *Learning and Individual Differences, 45*, 307–312. doi: 10.1016/j.lindif.2015.12.024
Pezzuti, L., Tommasi, M., Saggino, A., Dawe, J., & Lauriola, M. (2020). Gender differences and measurement bias in the assessment of adult intelligence: Evidence from the Italian WAIS-IV and WAIS-R standardizations. *Intelligence, 79*, 101436. doi: 10.1016/j.intell.2020.101436
Reynolds, M. R., Keith, T. Z., Ridley, K. P., & Patel, P. G. (2008). Sex differences in latent general and broad cognitive abilities for children and youth: Evidence from higher-order MG-MACS and MIMIC models. *Intelligence, 36*(3), 236–260. doi: 10.1016/j.intell.2007.06.003
Rodríguez-Cancino, M., & Concha-Salgado, A. (2023). WISC-V Measurement Invariance According to Sex and Age: Advancing the Understanding of Intergroup Differences in Cognitive Performance. *Journal of Intelligence, 11*(9), 180. doi: 10.3390/jintelligence11090180
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), 396–413. doi: 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. doi: 10.1111/j.1467-9280.1995.tb00301.x
Saggino, A., Pezzuti, L., Tommasi, M., Cianci, L., Colom, R., & Orsini, A. (2014). Null sex differences in general intelligence among elderly. *Personality and Individual Differences, 63*, 53–57. doi: 10.1016/j.paid.2014.01.047
Savalei, V., Brace, J. C., & Fouladi, R. T. (2023). We need to change how we compute RMSEA for nested model comparisons in structural equation modeling. *Psychological Methods*. doi: 10.31234/osf.io/wprg8
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*.
Scheiber, C. (2016a). Do the Kaufman tests of cognitive ability and academic achievement display construct bias across a representative sample of Black, Hispanic, and Caucasian school-age children in grades 1 through 12?. *Psychological Assessment, 28*(8), 942–952. doi: 10.1037/pas0000236
Scheiber, C. (2016b). Is the Cattell–Horn–Carroll-based factor structure of the Wechsler Intelligence Scale for Children—Fifth Edition (WISC-V) construct invariant for a representative sample of African–American, Hispanic, and Caucasian male and female students ages 6 to 16 years?. *Journal of Pediatric Neuropsychology, 2*, 79–88. doi: 10.1007/s40817-016-0019-7
Sipe, M. T. (2005). *Black-White differences in reading comprehension: The measure matters*. University of Maryland, College Park.
Sivo, S. A., Fan, X., Witta, E. L., & Willse, J. T. (2006). The search for “optimal” cutoff properties: Fit index criteria in structural equation modeling. *The Journal of Experimental Education, 74*(3), 267–288. doi: 10.3200/jexe.74.3.267-288
te Nijenhuis, J., Choi, Y. Y., van den Hoek, M., Valueva, E., & Lee, K. H. (2019). 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. doi: 10.1017/s0021932019000026
te Nijenhuis, J., & Van den Hoek, M. (2016). Spearman’s hypothesis tested on Black adults: A meta-analysis. *Journal of Intelligence, 4*(2), 6. doi: 10.3390/jintelligence4020006
te Nijenhuis, J., & Van Der Flier, H. (1997). Comparability of GATB scores for immigrants and majority group members: Some Dutch findings. *Journal of Applied Psychology, 82*(5), 675–687. doi: 10.1037/0021-9010.82.5.675
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. doi: 10.1016/s0160-2896(03)00027-8
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. doi: 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, second edition: A comparison among african american, asian, hispanic, and caucasian children. *Journal of Psychoeducational Assessment, 36*(7), 670-683. doi: 10.1177/0734282917698303

Van De Schoot, R., Schmidt, P., De Beuckelaer, A., Lek, K., & Zondervan-Zwijnenburg, M. (2015). Measurement invariance. *Frontiers in Psychology, 6*, 1064. doi: 10.3389/fpsyg.2015.01064
Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*(1), 4–70. doi: 10.1177/109442810031002
Van der Sluis, S., Derom, C., Thiery, E., Bartels, M., Polderman, T. J., Verhulst, F. C., ... & Posthuma, D. (2008). Sex differences on the WISC-R in Belgium and the Netherlands. *Intelligence, 36*(1), 48–67. doi: 10.1016/j.intell.2007.01.003
Van der Sluis, S., Posthuma, D., Dolan, C. V., de Geus, E. J., Colom, R., & Boomsma, D. I. (2006). Sex differences on the Dutch WAIS-III. *Intelligence, 34*(3), 273–289. doi: 10.1016/j.intell.2005.08.002
Walter, F., Daseking, M., & Pauls, F. (2021). Sex Differences in Intelligence in Children Aged 2: 6–7: 7: Analysis of the Factor Structure and Measurement Invariance of the German Wechsler Primary and Preschool Scale of Intelligence–Fourth Edition. *Journal of Psychoeducational Assessment, 39*(4), 395–421.
Warne, R. T. (2023). Tests of measurement invariance of three Wechsler intelligence tests in economically developing nations in South Asia and Sub-Saharan Africa. *Gifted and Talented International*, 1–17. doi: 10.1080/15332276.2023.2245007
Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: applications in the substance use domain. In K. J. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention* (pp. 281−324). Washington, DC: American Psychological Association. doi: 10.1037/10222-009
Wise, L. L., McLaughlin, D. H., & Steel, L. (1979). *The Project TALENT data handbook, revised*. Palo Alto, CA: American Institutes for Research.
Xiao, Y., Liu, H., & Hau, K. T. (2019). A comparison of CFA, ESEM, and BSEM in test structure analysis. *Structural Equation Modeling: A Multidisciplinary Journal, 26*(5), 665–677. doi: 10.1080/10705511.2018.1562928
Ximénez, C., Revuelta, J., & Castañeda, R. (2022). What are the consequences of ignoring cross-loadings in bifactor models? A simulation study assessing parameter recovery and sensitivity of goodness-of-fit indices. *Frontiers in Psychology, 13*, 923877. doi: 10.3389/fpsyg.2022.923877
Yoon, M., & Lai, M. H. (2018). Testing Factorial Invariance With Unbalanced Samples. *Structural Equation Modeling: A Multidisciplinary Journal, 25*(2), 201–213. doi: 10.1080/10705511.2017.1387859
Zhang, B., Luo, J., Sun, T., Cao, M., & Drasgow, F. (2023). Small but nontrivial: A comparison of six strategies to handle cross-loadings in bifactor predictive models. *Multivariate Behavioral Research, 58*(1), 115–132. doi: 10.1080/00273171.2021.1957664