Submitted: 9th of March 2025
Published: 30th of December 2025
DOI: 10.26775/OP.2025.12.30
ISSN: 2597-324X

Mixed evidence for Lynn's developmental theory of sex differences using aptitude tests

Meng Hu*

Abstract

This study investigates sex differences in the general factor of intelligence and their interaction with age, using Multiple Group Confirmatory Factor Analysis (MGCFA). It aims at testing Lynn's developmental theory of sex differences in intelligence, which states that the male advantage magnifies over the course of development, especially from age 16 onwards. The result provides some evidence for Lynn's hypothesis in the NLSY79 and NLSY97 but not in the Project Talent. Results from the Higher Order Factor (HOF) model showed that in the NLSY79, the male advantage in g increases from 1.21 to 5.53 points in the entire sample, while in the NLSY97, the male advantage increases from 0.18 to 2.46 points in the entire sample. Similarly, results from the Bifactor (BF) model showed a greater increase in g scores across ages among males. However, the BF model often produced substantially different score gaps in g in all three datasets. This discrepancy between the HOF and BF models highlights the influence of test composition on latent scores. A sibling pair analysis in the NLSY datasets yielded ambiguous results. In the Project Talent, sex differences remained stable across ages 14-18 in the White sample, but a slight increase in female advantage was observed in the Black sample, contradicting Lynn's hypothesis.

Keywords: Sex differences, aptitude tests, IQ, MGCFA, measurement invariance, Spearman's Hypothesis

1 Introduction

Tremendous efforts have been devoted to understanding the pattern of the sex differences in cognitive tests. A review literature by Reynolds et al. (2022) conclude that males have an advantage in visual processing whereas females have an advantage in processing speed, but they do not differ in general (g) factor scores. The last point is debatable because research sometimes found a g score advantage for men, and sometimes no sex differences.

Lynn's (2017, 2021) literature review shows that most studies on the Wechler's test reported a male advantage in full scale IQ, with this advantage being more pronounced and consistent in adult samples. Other cognitive tests also seem to support the conclusion of a male advantage in IQ. However, latent variable analyses of sex gaps on g are much less conclusive. It doesn't help that methods differ across studies, with Principal Component (PC), Multiple Indicator Multiple Causes (MIMIC), Multi-Group Confirmatory Factor Analysis (MGCFA), and which employ different competing models such as Higher Order Factor (HOF) of Bifactor (BF) model in the case of the MGCFA and the MIMIC. The evolutionary explanation for the male advantage in intelligence is that males compete for high status and for access to females, and the evolutionary explanation for the relatively greater male advantage in intelligence in adulthood is that early maturation affords a potential fitness advantage for females to begin reproducing in puberty as soon as they are sufficiently mature to have babies. Lynn finally argues that since there is a strong theoretical basis for the male advantage in IQ, one should not always blindly trust data: "As Einstein is said to have observed, "When the data and the theory are in conflict, it is generally the data that are wrong."." (p. 29).

*Independent researcher, Email: mh19870410@gmail.com

The issue of self-selection has been raised as a confounding factor in testing Lynn's hypothesis because women are more likely than men to participate voluntarily in surveys. Dykiert et al. (2009) argued that Lynn's developmental theory of sex difference is difficult to test properly due to selection bias that is more pronounced in adult samples, as children are easy to sample because all children must attend school while adults are more autonomous in their decisions: "Researchers must address the problem of non-inclusion of a proportion of their target population because there are systematic differences between those who choose to take part in studies and those who do not. These disparities may, and almost obviously do, bias the findings of studies on sex differences in intelligence, since it appears that both gender and intelligence may influence one's decision to participate." (p. 43). Their literature review, along with their analysis of the BCS, found evidence that attrition effects in follow-up waves caused a distortion in gender ratio, IQ means and variances, ultimately resulting in a higher female to male ratio, higher IQ for the remaining subjects, higher IQ variance for males, higher IQ means for males compared to females. Yet the sex IQ gap does not vary across ages, with a male advantage of 1.21, 1.79 and 1.39 IQ points at age 10, 26, and 30, respectively. A better method to estimate attrition is by way of logistic regression, proposed by Hunt & Madhyastha (2008), and which calculates the probability of participation in the follow-up depending on cognitive ability. Using this method, Madhyastha et al. (2009, Tables 3 & 6) analyzed the BCS and the NCDS data but found that the distortion, owing to differential attrition rates, on the male-female IQ is extremely small among adults. Even more troublesome is that Lynn & Kanazawa (2011, Tables 2-3) actually analyzed the NCDS among adolescents and found a very slight female advantage at age 7 and 11 but found a male advantage of 1.8 IQ points at age 16 as well as a consistently larger SD of males' IQ regardless of age. Because the pattern is similar whether the entire sample at each wave is analyzed or whether a restricted sample that completed all surveys is analyzed, they effectively ruled out the attrition effect as a possible explanation for the developmental theory of sex differences. Taken together, it would seem that the impact of differential attrition is unlikely to account for much of the observed gender full scale IQ difference, because this gap typically amounts to 5 points (Lynn & Irwing, 2004).

Perhaps, as a result, there is no definitive conclusion on sex differences in general ability. Differences in age, attrition effect and methodology across samples, however, do not explain why there is a general agreement in sex differences in specific abilities but not in general ability (Reynolds et al., 2022).

First, several studies report a female advantage in *g*. Reynolds et al. (2008) analyze the KABC-II, applying the MIMIC model to 5 different age groups (6-8, 9-11, 12-14, 15-16, 17-18) after establishing full measurement invariance with MGCFA. Based on the HOF model, the specification of *g* factor mean difference fixed to zero often did not worsen the model fit, which can be taken as evidence that the sex groups do not differ in *g*. And in a model that directly tests for age and sex interactions in the full sample, they found no age effects on sex gaps and report a 2-4 points advantage for females. Keith et al. (2008, Tables 6-9) analyze the Woodcock-Johnson III for individuals aged 6-59, using the MIMIC model for estimating sex and age interactions in *g*. They found a female advantage in *g* of 1.21 and 3.51 points based on the HOF and BF models, respectively, but their models could only assume partial invariance at the intercept level (i.e., subtest means). Out of 22 subtests, 7 showed intercept bias in the HOF model and 3 showed intercept bias in the BF model. Interestingly enough, they found an interaction between sex and age in the BF model only, which shows that the female advantage in *g* becomes larger over time, even reaching 5 IQ points at age 18-22. Härnqvist (1997) applied the BF model on a battery of 10 subtests in a sample of Swedish students from grade 4 to 9. These grade levels were analyzed simultaneously, revealing a modest advantage in *g* for females of about 3.45 IQ points for the entire sample. Rosén (1995, Table 6) analyzed a large battery composed of 13 cognitive tests and 15 achievement tests administered to 6th graders (ages 12-13) in Sweden. Results from the BF model, regardless of the full invariance or partial invariance specification, displayed a significant group difference favoring girls; the correlation between gender and latent *g* was 0.22, which produces a *d* gap of 0.45 (or 6.75 IQ points).[1] One could notice the results from these two independent studies conducted by Rosén and Härnqvist are not inconsistent with Lynn's prediction, given the young age of these samples.

But other studies report no sex gap in *g*. Keith et al. (2011, Table 6) applied the HOF model on the Differential Ability Scales among 5-17 years olds but found no sex differences in *g* and no interaction between sex and age. Lakin & Gambrell (2014, Table 9) compared sex differences in the CogAT, which is composed of 3 cognitive dimensions: sequential or deductive reasoning, quantitative reasoning, inductive reasoning. The battery measures primarily fluid ability (*gf*). Based on a BF model, there was no sex gap in *gf*, regardless of age groups (8-10, 11-13, and 14-17). Dolan et al. (2006) analyzed the WAIS-III in a spanish standardization sample and found no sex difference in *g*, based on the HOF model, even though measurement invariance was not tenable because no less than 4 subtests out of 14 were allowed to vary across groups due to intercept bias. Giofrè et al. (2024) tested

[1] This is based on r to Cohen's d conversion: $d = 2r/\sqrt{1-r^2}$ (Hunter & Schmidt, 2004, p. 279).

the HOF model in the Leiter-3, a nonverbal scale, on which they found 1 intercept bias and no statistically significant difference in the g factor mean although they did not report the magnitude and direction of the sex gap. This study probably had low statistical power due to the small sample size, given the difficulty to detect small differences in group means (Molenaar et al., 2009).

Finally, several studies report a male advantage in g. Lemos et al. (2013) compared sex differences among Portuguese students among young (13.5 yrs old) and old adolescents (16.9 yrs old), using a one-factor g model based on the 5 subtests of the Reasoning Test Battery. The males outperform females by 2 and 4 IQ points among young and old adolescents, respectively. Arribas-Aguila et al. (2019) applied a one-factor g model on the TEA Ability Battery composed of 7 subtests, 2 of which showed intercept bias and were therefore allowed to vary across gender. This model shows a null difference at ages 12-13 but a growing difference thereafter in favor of boys, with an advantage of 5 points in g scores at age 18. van der Sluis et al. (2006) analyzed adult Dutch twins in using MGCFA on the WAIS-III test. Based on the HOF model, one version of the g models fitted the data best, showing an advantage for men, but the authors did not report the effect size. van der Sluis et al. (2008) analyzed children twins from Netherlands and Belgium on the WISC-R test. They employed the same analysis and found a g advantage for males of 3.83 and 1.58 IQ points for the Dutch and Belgian samples, although they could not distinguish which models fit best between g models and non-g models and several subtests' means were allowed to vary across groups due to intercept bias. Irwing (2012) analyzed the WAIS-III using MGCFA, based on a Bifactor model, and reported a g advantage for men ($d = .22$) among adults aged from 16 to 89 years. Flores-Mendoza et al. (2013) administered both the Raven's SPM and the BPR5 battery (composed of 5 reasoning tests) on a sample of adolescent/adult Brazilians. A one-factor model of these 6 tests was fitted to the data, revealing a g advantage equivalent to 3.44 IQ points. Johnson & Bouchard (2007) analyzed the MISTRA data and applied the VPR model (a higher-order g model) on 42 subtests from 3 different cognitive batteries, and although they did not investigate measurement invariance, they reported a male advantage of $d = 0.14$ in g.

Amidst the inconsistent findings of sex differences, the issue of selection bias has been recognized by Deary et al. (2007) who argued that the solution for maximizing comparability is to compare opposite-sex, full siblings. This approach controls for factors such as SES, which differs between families but not within families. Upon analyzing the NLSY79 using this method, and based on the HOF model, they report a very small advantage in g for men ($d = 0.068$) at mean age 18.4.

Findings regarding to sex differences in g are ambiguous, just as the findings regarding Lynn's developmental theory at the g score level. Lynn (2017, 2021) presented results indicating that the sex IQ gap would appear at the age of 16 onwards, with an increasing male advantage. Using MIMIC, Keith et al. (2008, Table 9) found evidence that the score gap magnifies at age 16-17 onwards but in favor of females. Using MGCFA, Keith et al. (2011, Table 6) constrained age and age^2 effect on g to be equal across groups across ages 5-17 but found no degradation in model fit based on χ^2, indicating no age-by-sex interaction. They did not display the score gap by age group, especially toward age 16, where Lynn suggests the score gap will emerge. Using MIMIC, Reynolds et al. (2008, Figure 7) found no variation in the score gap across various age groups, specifically, 6-8, 9-11, 12-14, 15-16, 17-18.

The present study also uses latent variable methods (specifically, MGCFA) to uncover possible age-by-sex interaction in g among various aptitude tests. The purpose is to verify whether the magnitude of sex differences found in traditional IQ tests can be replicated in these aptitude tests, especially considering that traditional IQ tests are suspected to select test items in such a way that sex differences are minimized, or suppressed. One important caveat about aptitude tests is the sheer number of items requiring specific knowledge, on which males are typically more familiar with (Jensen, 1998, pp. 279, 534, 540). Despite these issues, aptitude tests are highly correlated with traditional IQ tests (Jensen, 1980, pp. 314-315, 330, 1998, p. 376) and have high predictive validity (Cucina et al., 2024; Hambrick et al., 2024; Jensen, 1980, pp. 347-353, 491-502; 1998, pp. 277, 286; Ree & Carretta, 2022).

2 Method

2.1 Data

The first dataset is the National Longitudinal Survey of Youth - 1979 Cohort (NLSY79), a longitudinal project that initially interviewed 12,686 American youth born between 1957-64. The respondents were aged 14-22 when first interviewed in 1979. The cognitive test battery, the Armed Services Vocational Aptitude Battery (ASVAB), was administered in 1981 and comprised 10 subtests that tap mostly general and technical knowledge: General Science (GS), Arithmetic Reasoning (AR), Word Knowledge (WK), Paragraph Comprehension (PC),

Numerical Operations (NO), Coding Speed (CS), Auto and Shop Information (ASI), Mathematics Knowledge (MK), Mechanical Comprehension (MC), Electronics Information (EI). A total of 11,914 participants completed the test. All analyses employ the sampling weights for the year 1981 (R0614600).

The second dataset is the National Longitudinal Survey of Youth - 1997 Cohort (NLSY97), a longitudinal project that initially interviewed 8,984 American youth born between 1980-84. The respondents were aged 12-17 when first interviewed in 1997. The cognitive test battery, the computer-adaptive form of the ASVAB (CAT-ASVAB), was administered in 1999 and comprised 12 subtests that tap mostly general and technical knowledge: General Science (GS), Arithmetic Reasoning (AR), Word Knowledge (WK), Paragraph Comprehension (PC), Numerical Operations (NO), Coding Speed (CS), Auto Information (AI), Shop Information (SI), Mathematics Knowledge (MK), Mechanical Comprehension (MC), Electronics Information (EI), Assembling Objects (AO). A total of 7,093 participants chose to participate in the CAT-ASVAB test. In the CAT form, the difficulty level is adaptive: if a respondent answers a question correctly (or incorrectly), the next question will be harder (easier). Among the advantages are the absence of mismatch between difficulty and examinee ability as well as the reduced risk of disclosure of testing material (Van der Linden & Glas, 2000). Despite the fact that respondents do not answer the same set of items, the scores are computed on a comparable scale using Item Response Theory.[2] All analyses employ the sampling weights for the year 1999 (R3923701).

A sibling analysis conducted by Deary et al. (2007) identified a very small sex IQ gap in the NLSY79. However, their analysis did not account for potential age-by-sex interactions, and the methodology for identifying opposite-sex full siblings was not clearly described. To address these gaps, the present study includes a sibling analysis of both the NLSY79 and NLSY97 datasets. Full siblings were identified based on a sibling genetic relatedness coefficient of 0.5 using the *NlsyLinks* package (Beasley et al., 2024). Specifically, the *Links79Pair* and *Links97Pair* datasets were linked to the NLSY79 and NLSY97 datasets, respectively. Households with only one member were excluded from the analysis. Similarly, households without opposite-sex siblings were removed. For households with more than two opposite-sex siblings, an imbalance in the ratio of male to female siblings was often observed. This imbalance was particularly pronounced in families with an odd total number of siblings (e.g., 3, 5, or 7 siblings) and was also occasionally present in families with an even total number of siblings (e.g., 4 or 6 siblings). To ensure balanced representation, younger siblings were iteratively excluded within each sex category. This process continued until all households included an equal number of male and female siblings (e.g., two brothers and two sisters in a household with four siblings).

The third dataset is the Project Talent, 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 employ the student weight variable (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. A prior study examined the gender differences in g for this dataset, but did not test Lynn's developmental hypothesis (Hu, 2025).

A total of 34 subtests are used in this analysis, following Hu (2025). They represent 6 different cognitive/aptitude areas such as: 1) English factor (disguised words, spelling, capitalization, punctuation, English use, effective expression, word function, reading comprehension), 2) math factor (math, arithmetic reasoning, high school math, arithmetic computation), 3) processing speed factor (table reading, clerical checking, object inspection), 4) knowledge/information factor (vocabulary, literature, music, social science, art, law, health, bible, theater, miscellaneous), 5) science knowledge factor (physical science, biological science, aeronautics and space, electronics, mechanics) 6) spatial reasoning factor (mechanical reasoning, visualization in 2D, visualization in 3D, abstract reasoning). As a robustness test, MGCFA was also carried out after removing the subtests that have their main loading on the information factor and science factor to account for Jensen's criticism (1985, p. 218). Several subtests loaded on multiple factors, and these cross loadings were specified following cutoff recommendations from simulation studies suggesting that small/modest loadings (e.g., .15-.20) can be included (Cao & Liang, 2023; Hsu et al., 2014; Xiao et al., 2019; Ximénez et al., 2022; Zhang et al., 2023).

[2] A description of the procedure is provided here: https://www.nlsinfo.org/content/cohorts/nlsy97/other-documentation/codebook-supplement/appendix-10-cat-asvab-scores

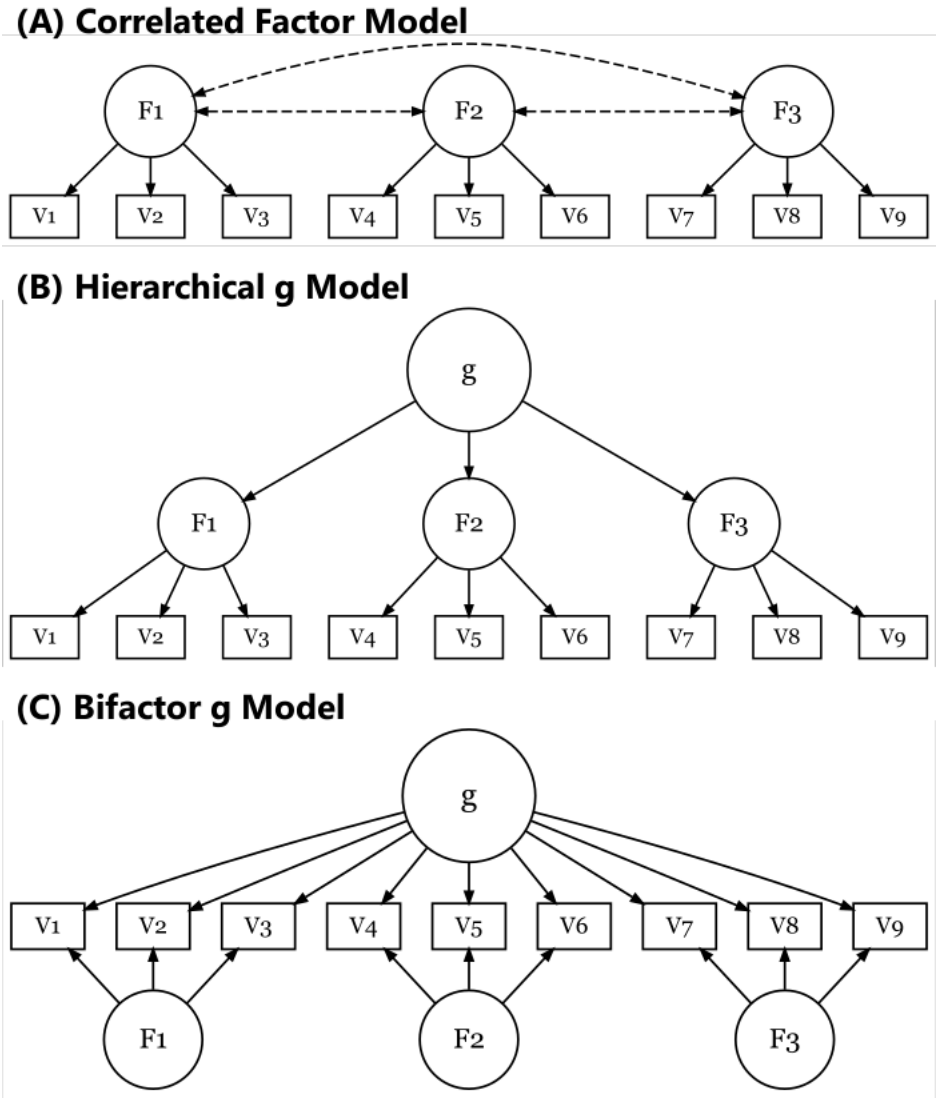

Figure 1: Illustration of the competing CFA models.

2.2 Statistical model

All statistical analyses are done using R and, in particular, the *lavaan* package for MGCFA models. The goal is to evaluate the magnitude of the sex difference in g across different g models, as well as the Spearman's hypothesis, which states that the observed group differences in the (sub)tests are primarily due to g. 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. Figure 1 displays hypothetical competing CFA models that are investigated in the present analysis:

The CF model assumes that the group (i.e., specific) factors are correlated without even specifying a general factor; the HOF model assumes that the relationship between the general factor and subtests is mediated by the first-order specific factors; the BF model assumes that both the general and specific factors are uncorrelated and have a direct relationship with the subtests. If the model fit of either g model shows superior fit compared to the non-g model, it suggests that g represents the data better than a model without g. Finally, the best g model can be used to test Spearman's hypothesis, by calculating the proportion of the subtests' group differences that are due to g and non-g factors.

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 χ^2. Higher values of CFI and Mc indicate better fit, while lower values of χ^2, RMSEA, RMSEA_D, SRMR indicate better fit. Simulation studies have 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) as well as Zhou (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.

Several 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 ΔCFI 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 ΔMc .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 $\geq$.005 in CFI, supplemented by a change of $\geq$.010 in RMSEA or a change of $\geq$.025 in SRMR; for testing intercept or residual invariance, a change of $\geq$.005 in CFI, supplemented by a change of $\geq$.010 in RMSEA or a change of $\geq$.005 in SRMR. The values of ΔMc 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.

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.[3] 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.

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 ($\nu_1 \neq \nu_2$), 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

[3] 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 (Hu, 2025). This strict invariance assumption is however useful when the focus analysis is the reliability of the construct the latent variable represents (Beaujean, 2014, p. 59).

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[4]

Model	Specification	Nesting
M1. Configural	$\lambda_1 \neq \lambda_2 + \Psi_1 \neq \Psi_2 + \nu_1 \neq \nu_2 + \Theta_1 \neq \Theta_2 + \delta = 0$	
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) $\nu_1 = \nu_2$ + (all) $\delta \neq 0$	under M2a
M3a. Partial Scalar	M2a but adds (partial) $\nu_1 = \nu_2$ + (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) $\delta_{\text{non-g}} = 0$	under M5
M6b. Weak SH	M5 but adds (partial) $\delta_{\text{non-g}} = 0$	under M5
M6c. No SH	M5 but adds $\delta_g = 0$	under M5

Note: λ = loadings; Ψ = covariance; ν = intercepts; Θ = residuals; δ = factor means.

Insofar as the study of sex differences relates to Spearman's g, the most appropriate latent model is the Bifactor. As Beaujean (2015) observed, "Spearman believed that g was directly involved in all cognitive "performances", not indirectly involved through, or mediated by, other factors." (p. 130). According to Dolan (2000), an unambiguous test of the Spearman's Hypothesis should compare different competing models, including models that do not include the g factor. For this reason, three competing models were fitted to all data: the Correlated-Factors (CF) model, the Higher-Order-Factor (HOF) model, and the Bifactor (BF) model.

A test of Lynn's hypothesis involves a model that includes the interaction between age and sex. A latent MIMIC model would be ideal for this purpose, but the MIMIC typically assumes group equality of factor loadings, error variances and factor variances. Because both error variances and factor variances are violated in the present analysis, this method is not employed. Instead, the g factor from the best HOF/BF model is regressed on age, providing separate age regression coefficients for males and females.

2.3 Data preparation and assumption tests

For both the NLSY79 and NLSY97, there was no need to employ multiple imputation because there is no missing data in any of the subtests. Univariate normality is assessed, and established with cutoffs of 2.0 for skewness and of 7.0 for kurtosis (Curran et al., 1996). Multivariate normality was assessed using the mqqnorm() function of the *stats* package and the mvn() function from the *MVN* package; the results show that the deviation from normality was small for NLSY79 but large for NLSY97. Exploratory Factor Analysis (EFA) determined that the 4 factor model was the most interpretable for both datasets. The usage of sampling weight in CFA modeling requires robust maximum likelihood estimation, which has the peculiarity to be robust to non-normality.

Regarding the Project Talent, a prior study (Hu, 2025) showed that multivariate normality was not tenable, but this issue can be circumvented by the use of Robust Maximum Likelihood estimation and that the Exploratory Factor Analysis (EFA) suggested that the 6-factor model was the most interpretable in all subgroups tested. In this study, small loadings (.15-.20) were used because several simulations suggested that loadings of this magnitude improves the sensitivity of fit indices. However, in the case of latent factor modeling, the cross loadings are often found to be much smaller in the BF model compared to other factor models such as CF and HOF. Thus, for BF models, cross loadings lower than ≈.10 are removed but retained when greater than ≈.10. This is because with such a large model and sample size, these loadings are still significant and their inclusion also improves the model fit.

[4] In the CF model, the factor variance as well as the covariance among factors eventually need to be constrained to be equal.

2.4 Meta-analysis of sex differences in IQ test batteries

To shed light on the huge heterogeneity of effect sizes, a meta-analysis was performed to assess whether this heterogeneity follows a discernible pattern. The data are restricted to studies reviewed in the Introduction section, as they use comprehensive IQ test batteries (both the ASVAB and the Project Talent batteries are best characterized as aptitude tests; while excellent proxies for IQ, they comprise numerous subtests that favor males). The literature search relied primarily on Google Scholar, with supplementary searches conducted in PubMed and PsycINFO.

In meta-analysis, both the effect size and standard errors must be computed following strict criteria. As for the effect size, often Cohen's d was not reported, and instead the correlation or the IQ-metric effect size was provided. In these cases, r was converted to d following the formula provided by Hunter & Schmidt (2004, p. 279) or the IQ-metric difference was converted to d-metric difference. As for the standard errors, they were often not reported, so these values were computed based on the standard formula based on sample size provided by Hunter & Schmidt (2004, p. 284).

The moderators examined were age, number of subtests, and factor model type (HOF or BF). When a mean age was not reported, the median of the provided age range was used. Studies which use a one-factor model (due to the small number of subtests) were classified as HOF, because these models typically yield similar estimates of g. This would allow one to examine the difference in effect size between HOF and BF models.

A classic meta-analysis may be inappropriate however, due to within-study dependencies. Specifically, two studies (Arribas-Aguila et al., 2019; Reynolds et al., 2008) report effects for different age cohorts, while one study (van der Sluis et al., 2008) used two independent samples, which we treated as separate effect sizes. To avoid such within-study dependencies, a three-level multilevel approach which accounts for this nested structure was undertaken for the meta-analytic model (Harrer et al., 2021). A more complex modeling such as the Correlated Hierarchical Effects (CHE) is not needed because the within-study effects came from distinct samples, as only Keith et al. (2008) study displayed correlated effects. Thus a traditional three-level multilevel model is sufficient for the present analysis.

3 Result

3.1 NLSY79 & NLSY97

For the NLSY79, for all study samples, the model specification is displayed as follows:

$$\begin{aligned} \text{verbal} &\sim \text{GS} + \text{WK} + \text{PC} \\ \text{math} &\sim \text{AR} + \text{MK} + \text{MC} \\ \text{electronic} &\sim \text{GS} + \text{ASI} + \text{MC} + \text{EI} \\ \text{speed} &\sim \text{NO} + \text{CS} \end{aligned}$$

The model has to be modified for the bifactor specification, because the loading of WK on verbal was out of bound or the variance of WK was out of bound, probably because its loading on g were high ($>.90$). Even after removing this loading, the verbal factor was not clearly identified because the loading of GS on verbal was strongly negative. This indicates that verbal is redundant once g is accounted for, consistent with the very high loading ($>.95$) of verbal factor on g in the HOF models. Thus, the verbal factor was removed.

For the NLSY97, for all study samples, the model specification is displayed as follows:

$$\begin{aligned} \text{verbal} &\sim \text{GS} + \text{WK} + \text{PC} + \text{EI} \\ \text{math} &\sim \text{AR} + \text{MK} + \text{MC} + \text{AO} \\ \text{electronic} &\sim \text{AI} + \text{SI} + \text{MC} + \text{EI} \\ \text{speed} &\sim \text{NO} + \text{CS} + \text{MK} \end{aligned}$$

Once more, the bifactor specification required some adjustments. In the White sample, the math factor was ill defined due to abnormal values in the variance and loading of AO, and removing its loading on math caused

the loading of MC on math to be negative. Thus, the math factor was omitted. Due to potential criticism and concern about the changing nature of g (Eid et al., 2017; Eid, 2020; Koch & Eid, 2024), a model which uses the same specification as in other samples (i.e., omitting verbal factor) has also been fitted to the data.[5] The fit of this model was actually worse (see supplementals). In the all-race sample, the verbal factor was ill defined because only WK has high loading and EI has a near-zero loading. Removing EI causes WK to be out of bound, and removing WK causes GS to have negative loading. Thus, the verbal factor was omitted.

3.1.1 NLSY79

Analyses revealed strong measurement bias at the intercept level for both the CF and HOF models in both the non-Hispanic White and all-race samples of the NLSY79. Results are discussed and displayed in the appendix (Tables A1-A2). In all cases, the most parsimonious model was one which allowed partial invariance in the intercepts and latent factor variances. In this model, the estimated male advantage was as follows: in the non-Hispanic White sample, 4.5 IQ units for the HOF model and -0.5 IQ units for the BF model; in the all-race sample, 3.2 IQ units for the HOF model and -0.4 IQ units for the BF model. The standardized sex differences in latent means for these models are displayed in Table 2.

Table 2: Standardized sex difference in latent means in the NLSY79 for best HOF/BF models

	White sample		All-race sample	
Factor	**HOF**	**BF**	**HOF**	**BF**
Verbal	.578 (.027)	(omitted)	.438 (.019)	(omitted)
Math	0 (fixed)	−.509 (.039)	0 (fixed)	−.467 (.032)
Electronic	−1.441 (.054)	−1.915 (.060)	−.969 (.029)	−1.670 (.042)
Speed	.516 (.028)	.604 (.047)	.430 (.021)	.588 (.038)
g	−.298 (.035)	.034 (.031)	−.214 (.028)	.024 (.023)

Note: Standard errors in parentheses. Negative values denote a male advantage.

To evaluate Lynn's theory, the g factor was regressed on age using the final HOF and BF models. This analysis was conducted separately for the non-Hispanic White sample (Table A1) and the all-race sample (Table A2). In both analyses, imposing a group equality constraint on the age regression coefficient slightly deteriorated the model fit (based on SRMR), which was unsurprising given the slight differences in point estimates between groups. The results showed a consistent pattern across samples. For the non-Hispanic White sample, the HOF model indicated a male advantage that increased with age from 2.02 IQ points at age 14 to 7.18 IQ points at age 22. The BF model for this sample, however, showed a female advantage of 2.44 points at age 14, which shifted to a male advantage of 1.63 points by age 22. In the all-race sample, this pattern was mirrored, as the HOF model showed a male advantage increasing from 1.21 to 5.53 IQ points. Similarly, the BF model for this sample showed a female advantage of 2.08 points at age 14 that also shifted to a male advantage of 1.63 points by age 22.

Additional analyses are also carried out to verify whether the pattern of sex differences found in the main analyses replicated using the opposite-sex sibling pair approach and cross validation samples. In the sibling data, rather surprisingly, a series of MGCFA models reveal a consistent female advantage in the growth of g scores, as indicated by the regression coefficient of age. This pattern holds across both HOF and BF models and in both the White and all-race samples. Overall, males show an advantage of approximately 0.7 IQ point, except in the HOF model for the all-race sample, where the male advantage is larger at 2.7 points. The results are cross validated using a split-half approach which is based on random half samples. 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 .003, except some specification in the CF model for the White sample. More importantly, the original constraints hold just as well in each random half sample.

3.2 NLSY97

Analyses revealed strong measurement bias at the intercept level for all 3 models (CF, HOF, BF) in both the non-Hispanic White and all-race samples of the NLSY97. Results are discussed and displayed in the appendix

[5] Although the source of the misfit is related to the math subtests, omitting the verbal factor instead of the math factor also solved the inadmissible solution, but at the cost of worse fit.

(Tables A3-A4). In all cases, the most parsimonious model was one which allowed partial invariance in the intercepts and latent factor variances. In this model, the estimated male advantage was as follows: in the non-Hispanic White sample, 1.1 IQ units for the HOF model and -2.2 IQ units for the BF model; in the all-race sample, 1.2 IQ units for the HOF model and 2.5 IQ units for the BF model. The standardized sex differences in latent means for these models are displayed in Table 3.

Table 3: Standardized sex difference in latent means in the NLSY97 for best HOF/BF models

	White sample		All-race sample	
Factor	**HOF**	**BF**	**HOF**	**BF**
Verbal	.052 (.011)	−.321 (.056)	0 (fixed)	(omitted)
Math	.050 (.021)	(omitted)	.068 (.017)	.356 (.044)
Electronic	−.770 (.028)	−1.131 (.043)	−.596 (.020)	−.744 (.028)
Speed	.530 (.036)	0 (fixed)	.519 (.028)	.902 (.054)
g	−.076 (.034)	.150 (.036)	−.080 (.026)	−.167 (.027)

Note: Standard errors in parentheses. Negative values denote a male advantage. In the White sample, for the BF model, if verbal is omitted instead of the math factor, the means (standard errors) are .511 (.068), −.828 (.037), 1.066 (.078), −.217 (.036) for math, electronic, speed and *g*, respectively.

One issue with the BF model specification is the changing nature of *g*. By omitting math instead of verbal, the *g* score in the White sample reflects primarily math instead of verbal. According to Eid et al. (2017), the omission of a specific factor should be theory-driven rather than data-driven. However, in all samples, it was observed in the HOF model that both verbal and math factors had a strong loading (>.90) on *g*, even though the loading of verbal on *g* was greater than the loading of math on *g*. It is therefore not clear that *g* in the ASVAB reflects primarily verbal ability. A more accurate depiction is that ASVAB reflects mainly verbal as well as math abilities. Nonetheless, a BF model which omits the verbal factor has been fitted to the data. Compared to the BF model that omits the math factor, it displayed a worse fit (CFI = .962 instead of CFI = .966) and a strikingly different gender gap in *g* (3.25 IQ points in favor of male instead of 2.25 IQ points in favor of females), as shown in Table 2.

To evaluate Lynn's theory, the *g* factor was regressed on age using the final HOF and BF models. This analysis was conducted separately for the non-Hispanic White sample (Table A3) and the all-race sample (Table A4). In both samples, imposing a group equality constraint on the age regression coefficient did not affect model fit (based on SRMR), which was unsurprising given that the point estimates for men and women differed only slightly. The results, however, varied by sample and model. For the non-Hispanic White sample, the HOF model showed a small male advantage that increased with age, from 0.34 IQ points at age 12.4 to 2.32 IQ points at age 16.4. In contrast, the BF model for this sample indicated a female advantage of 3.06 and 1.02 IQ points at ages 12.4 and 16.4, respectively. When the verbal factor was omitted from the BF model, this pattern reversed, showing a male advantage increasing from 2.70 to 4.32 IQ points. In the all-race sample, both models showed a male advantage that increased with age. The HOF model yielded an advantage of 0.18 IQ points at age 12.4, rising to 2.46 IQ points at age 16.4. The BF model showed a larger increasing advantage, from 1.45 to 3.79 IQ points across the same ages.

Additional analyses are also carried out to verify whether the pattern of sex differences found in the main analyses replicated using the opposite-sex sibling pair approach and cross validation samples. In the sibling data, a series of MGCFA models reveal a consistent male advantage in the growth of *g* scores, as indicated by the regression coefficient of age. This pattern holds across both HOF and BF models and in both the White and all-race samples. Overall, males show an advantage of approximately 1 IQ point, except in the BF model for the White sample, where the male advantage is notably larger at 6 points. The main analyses are cross validated using a split-half approach which is based on random half samples. 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 .003 in the all race sample, but can vary between .007 and .010 in some specification for either the CF, HOF, or BF model for the White sample. More importantly, the original constraints hold just as well in each random half sample.

3.3 Project Talent

3.3.1 MGCFA: White sample

The model specification for this sample is displayed as follows:

$$\begin{aligned}
\text{english} \sim\ & \text{DWords} + \text{Spelling} + \text{Capitalization} + \text{Punctuation} + \text{EnglishUse} + \text{Expression} \\
& + \text{WordF} + \text{ReadComp} + \text{AbsReason} + \text{ArithReason} + \text{ArithComp} \\
\text{math} \sim\ & \text{Math} + \text{PhySci} + \text{WordF} + \text{ArithReason} + \text{HSMath} + \text{ArithComp} \\
\text{speed} \sim\ & \text{DWords} + \text{Visual2D} + \text{ArithComp} + \text{TableRead} + \text{ClericalCheck} + \text{ObjectInspect} \\
\text{info} \sim\ & \text{Vocab} + \text{Literature} + \text{Music} + \text{SocialSci} + \text{BioSci} + \text{AeroSpace} + \text{Art} + \text{Law} \\
& + \text{Health} + \text{Bible} + \text{Theater} + \text{Misc} + \text{DWords} + \text{ReadComp} \\
\text{science} \sim\ & \text{Vocab} + \text{PhySci} + \text{BioSci} + \text{AeroSpace} + \text{Elec} + \text{Mechanics} + \text{Health} + \text{MechReason} \\
\text{spatial} \sim\ & \text{MechReason} + \text{Visual2D} + \text{Visual3D} + \text{AbsReason} + \text{ArithReason} + \text{ObjectInspect}
\end{aligned}$$

Analyses revealed strong measurement bias at the intercept level for all 3 models (CF, HOF, BF) in the White sample. Results are discussed and displayed in the appendix (Table A5). In all cases, the most parsimonious model was one which allowed partial invariance in the intercepts. In this model, there was a large male advantage of 5.4 and 14 IQ units in the HOF and BF models, respectively. The standardized sex differences in latent means for these models are displayed in Table 4. Lynn's hypothesis failed in this sample, because the regression of g on age showed that the regression coefficients of age were almost identical across sexes in both the HOF and BF models.

Table 4: Standardized sex difference in latent means in the Project Talent for best HOF/BF models, using the set of 34 tests (White sample)

Factor	HOF	BF
English	.966 (.005)	2.946 (.030)
Math	0 (fixed)	.871 (.023)
Speed	.421 (.007)	.543 (.008)
Information	.363 (.005)	2.313 (.024)
Science	−1.904 (.012)	−2.238 (.022)
Spatial	−.522 (.009)	−.202 (.014)
g	−.360 (.007)	−.931 (.014)

Note: Standard errors in parentheses. Negative values denote a male advantage.

3.3.2 3.2.2 MGCFA: Black sample

The model specification for this sample is displayed as follows:

$$\begin{aligned}
\text{english} \sim\ & \text{Vocab} + \text{DWords} + \text{Spelling} + \text{Capitalization} + \text{Punctuation} + \text{EnglishUse} + \text{Expression} \\
& + \text{WordF} + \text{ReadComp} + \text{AbsReason} + \text{ArithComp} \\
\text{math} \sim\ & \text{Math} + \text{WordF} + \text{ArithReason} + \text{HSMath} + \text{ArithComp} \\
\text{speed} \sim\ & \text{DWords} + \text{ArithComp} + \text{TableRead} + \text{ClericalCheck} + \text{ObjectInspect} \\
\text{info} \sim\ & \text{Vocab} + \text{Literature} + \text{Music} + \text{SocialSci} + \text{BioSci} + \text{AeroSpace} + \text{Art} + \text{Law} \\
& + \text{Health} + \text{Bible} + \text{Theater} + \text{Misc} + \text{DWords} + \text{ReadComp} \\
\text{science} \sim\ & \text{Vocab} + \text{Math} + \text{PhySci} + \text{BioSci} + \text{Elec} + \text{Mechanics} \\
\text{spatial} \sim\ & \text{MechReason} + \text{Visual2D} + \text{Visual3D} + \text{AbsReason} + \text{ObjectInspect}
\end{aligned}$$

Analyses revealed strong measurement bias at the intercept level for all 3 models (CF, HOF, BF) in the Black sample. Results are discussed and displayed in the appendix (Table A6). In all cases, the most parsimonious model was one which allowed partial invariance in the loadings and intercepts. In this model, there was a large male advantage of 2.3 and 7.2 IQ units in the HOF and BF models, respectively. The standardized sex differences in latent means for these models are displayed in Table 5. Lynn's hypothesis failed in this sample, because the regression of g on age showed that the regression coefficients of age were similar across sexes despite the coefficient being slightly larger for females in both the HOF and BF models.

Table 5: Standardized sex difference in latent means in the Project Talent for best HOF/BF models, using the set of 34 tests (Black sample)

Factor	**HOF**	**BF**
English	.516 (.026)	1.707 (.098)
Math	0 (fixed)	.569 (.104)
Speed	.288 (.034)	.428 (.038)
Information	.161 (.026)	1.203 (.103)
Science	−.581 (.039)	−1.505 (.085)
Spatial	−.448 (.036)	−.301 (.056)
g	−.153 (.036)	−.481 (.054)

Note: Standard errors in parentheses. Negative values denote a male advantage.

3.3.3 MGCFA: Robustness test

The findings of very large gender g gaps are consistent with a previous study (Hu, 2025) but not necessarily surprising. As Jensen (1998, pp. 279, 534, 540) mentioned, there is indeed a large male advantage in specific knowledge elicited by the tests of scientific knowledge and social science. And this was indeed observed in the data. For this reason, MGCFA was reconducted after removing these subtests. For both the White and Black samples, the exploratory factor analysis revealed that the 4-factor solution yielded the most interpretive and clearest pattern (compared to 3- and 5-factor solutions), namely, english, math, speed and spatial factors. The standardized sex differences in latent means are displayed in Table 6.

Table 6: Standardized sex difference in latent means in the Project Talent for best HOF/BF models, using the set of 20 tests

	White sample		Black sample	
Factor	**HOF**	**BF**	**HOF**	**BF**
English	.721 (.006)	2.214 (.017)	.397 (.017)	.854 (.060)
Math	−.428 (.006)	.086 (.009)	−.126 (.014)	−.473 (.066)
Speed	.336 (.007)	.422 (.007)	.276 (.033)	.282 (.036)
Spatial	−.394 (.007)	−.218 (.009)	−.479 (.030)	−.768 (.046)
g	−.050 (.008)	−.352 (.007)	−.082 (.030)	.066 (.031)

Note: Standard errors in parentheses. Negative values denote a male advantage.

Results from MGCFA in the White sample showed that the BF model fitted much better than the CF model. In the BF model, when iteratively freeing intercepts, it was noticed that the sex difference in g changes modestly. For instance, the last intercept to be freed for the model fit to be acceptable was either math or reading comprehension, both of which exhibited very similar chi-square values. In the final model, which sets a group equality constraint on all factor variances, there is a male advantage in g of .352 when reading comprehension was released and .281 when math was released. Deciding which subtest's intercept should be released was no easy task because both had a very strong loading on g. But it was decided to release the constraint on reading comprehension because its loading on the english factor was small. The reason why the impact of partial scalar invariance on g differences is larger than in the previous model that uses a larger number of subtests is because the individual impact of each subtest on g or specific factors is greater as the total number of subtests decreases.

Results from MGCFA in the Black sample showed that the BF model fitted much better than the CF model. In the BF model, only one intercept (namely, vocabulary) had to be freed, and it led to a very small change in g. In the final model, which sets a group equality constraint on all factor variances, there is a female advantage in g of .068. Here again, as in the White sample, the sex difference in g is much smaller in this reduced battery. This shows that the type of test contained in the battery affects the magnitude of the sex difference in g.

3.4 Multilevel meta-analysis of sex IQ differences

The meta analysis is based on 25 effects. Initially, a model without moderator was fitted to the data, revealing a small, non-significant male advantage ($d = 0.019$, S.E. $= 0.059$, p $= 0.749$). Then, a model including all three moderators was fitted, but it produced a suspicious between-study variance estimate of zero, suggesting overfitting. As the number of subtests exhibited a very weak coefficient and was the least theoretically justified moderator, it was removed. The final meta-analysis was conducted using only age and factor model as moderators. Results, displayed in Table 7, showed two interesting outcomes. First, age had a weak impact on sex differences, yet the positive sign indicates that the male advantage increases with age. Second, factor model had a sizable impact on sex differences in g, with the HOF model exhibiting a larger male advantage ($d = .247$) compared to the BF model. Such a straightforward interpretation would be overly simplistic however. As can be observed in the present analysis of the ASVAB and Project Talent batteries as well as in the literature review (e.g., Keith et al., 2008), BF models have a tendency to produce effect sizes larger than HOF models, regardless of the direction of the sex differences.

Table 7: Parameter estimates of the multilevel meta-analysis

	Estimate	**S.E.**	**p-value**	**C.I.**
Intercept	−.307	.089	.002	[−.492 : −.122]
Age	.008	.003	.021	[.001 : .014]
Model HOF	.247	.072	.002	[.097 : .396]

Variance components analysis indicated substantial true heterogeneity ($I^2 = 88.1\%$), with approximately half arising from between-study methodological differences (I^2 level 3 $= 54.2\%$) and one-third from within-study age effects (I^2 level 2 $= 33.9\%$). The remaining 11.9% represented sampling error. Finally, funnel plot analyses reveal no publication bias.

4 Discussion

The present study showed that Lynn's developmental theory is borne out by the NLSY datasets but not the Project Talent dataset. In the NLSY79, there is a growth of the male advantage in g by about 4-5 points in the White sample and 4 points in the all-race sample across ages 14-22 years. In the NLSY97, there is a growth of the male advantage in g by about 2 points in the White sample and 2 points in the all-race sample across ages 12.4-16.4 years. In the Project Talent, there is no gender difference in cognitive growth based on a large set of 34 subtests or even a reduced set of 20 subtests (that excludes the tests of knowledge). The magnitude of the average sex differences also varies across samples, models (HOF vs BF) and test composition.[6]

There are some difficulties in interpreting these results because scalar measurement invariance did not hold for both the NLSY79/97 and Project Talent and because the composition of the test batteries affected the magnitude of the gender difference in g and non-g factor means. The impact of partial scalar invariance on these latent means depends on which subtest is set free, and the impact is stronger among non-g factors and for smaller test batteries.[7] For data that has non-invariant parameters, it is still recommended to employ partial invariance because it recovers the true values of the group means and regression paths much better than the full invariance (Pokropek et al., 2019). Moreover, the observation that higher-order factor g and bifactor g models

[6] A reviewer conducted a meta-analysis of the 16 effect sizes obtained in this study to synthesize the overall sex difference. The simple random-effects model indicated a small, significant average male advantage (estimate = 0.19, p = .005). However, there was significant and substantial heterogeneity ($I^2 = 99.63\%$), suggesting that the effect size varied substantially depending on the specific test battery, sample, or measurement model used. The full analysis is available at: https://rpubs.com/EmilOWK/meng_hu_2025_sex_diffs_g

[7] In one case however, the impact of partial scalar on g was large. In the White sample of the NLSY97, based on the BF model which omits verbal factor, the full scalar and partial scalar models displayed a male advantage of 0.34 and 3.25 points, respectively. This change was the result of releasing the constraints on two verbal subtests, namely PC and WK.

produce very different estimates of sex differences in g implies that test composition matters for the study of sex differences. This might seem surprising given that earlier research has established that the nature of the g factor is robust to test composition, concluding that there is only one g (Johnson et al., 2004, 2008). Regarding this point, there are two observations to be made: first, these studies only employed and compared higher-order factor g models and, secondly, the mean structure and correlation structure are completely unrelated (see, e.g., Rodgers, 1998, p. 343).

Due to the presence of scalar non-invariance and its impact on latent means, a direct test of the Spearman's hypothesis becomes less informative as it requires a decomposition analysis to separate the percentage of the subtest difference due to g and non-g factors (Dolan, 2000; Hu, 2025). This analysis was not attempted here because in the NLSY79/97, the sex difference in g often vanishes or reverses in the BF model, and in the Project Talent, such a test was already carried out (Hu, 2025).

Overall, the present finding confirms Jensen's (1998) suspicions about aptitude tests favoring males due to the sheer number of tests related to scientific knowledge and social science. There are still expressed concerns about test content playing a modest role in shaping the direction and magnitude of the sex gap. Lemos et al. (2013) speculated that "in early elementary school years, quantitative ability is measured mainly through computational tasks, at which girls outperform boys; as we go through higher secondary-school grades, mathematical concepts require more reasoning abilities and become more spatial in nature (e.g. problem solving in geometry and calculus), which favors boys." (p. 12). In their meta-analysis, Giofrè et al. (2022) observed that while the full scale IQ shows a male advantage of 1.395 IQ points in the WISC, this advantage drops to 0.81 IQ points when only the newer version of the WISC is considered. Moreover, Piffer (2016) analyzed the magnitude of the full scale and subtest IQ gaps in the WAIS-IV across gender and concluded that the patterns are similar to the results obtained in previous standardization samples of the WAIS, WAIS-R, and the WAIS-III despite changes in content across those versions. As Piffer (2016) noted, this happened despite the fact that test developers continuously attempted to remove items displaying observed sex differences. However, these studies still did not evaluate g. Why this matters is best illustrated by the findings from Keith et al. (2008) who reported that the sex difference in g is 2 points larger in the BF model compared to the HOF model in the Woodcock-Johnson III. Thus, even in traditional IQ batteries, test content affects the nature of specific factors, which may create those discrepancies in g scores that we might observe between HOF and BF models.

It might be argued, given the present result, that the bifactor structure does not represent the data well despite superior model fit. In the NLSY79 and NLSY97, the bifactor model often results in one factor—typically verbal—disappearing. This, along with the presence of both low loadings on specific factors and high loadings on the g factor in both the NLSY and Project Talent datasets, raises legitimate concerns about the validity of a bifactor structure (Eid et al., 2017; Watts et al., 2019). On the other hand, DeMars (2013) noted, in the case of the bifactor model, that "When loadings on the specific factors are low, only the general factor score carries a reliable interpretation." This is particularly insightful given that this study's focus is on the g factor score. More importantly, Beaujean et al. (2014) did not seem to consider a vanishing factor as a reason to reject the bifactor structure. In their study, the fluid reasoning factor was removed in the bifactor model because it was not necessary to explain the covariances. This was because fluid and g were indistinguishable, an outcome that was confirmed in their higher order factor model as well. In the bifactor, fluid becomes redundant once g is accounted for. In the case of the NLSY79/97, the omission of a verbal factor can be taken as an indication that g and verbal factors are almost indistinguishable.[8] This is somewhat consistent with Roberts' et al. (2001) conclusion that the ASVAB has a very strong verbal flavour. However in the NLSY97, both the math and verbal factors had extremely high loading on g, and in the White sample, a specification omitting math factor (instead of verbal) fitted the data better but also changed the nature of g and caused the sex score gap to reverse (see Table 3). This adds to the challenge of interpreting bifactor results when measurement invariance is strongly violated, when partial invariance changes the magnitude of g scores and when the loadings of multiple factors on g are so high. This does not mean the bifactor g model should be discarded. Instead, this means the NLSY data is not ideal for testing the bifactor structure.

Another limitation of the present study is the absence of group equality on g even when the group difference was non-significant in the final BF models for the NLSY79 (see Table 2). This was justified because the question that needs to be answered is whether g scores interact with age in a way consistent with Lynn's prediction. Imposing a group equality constraint on g would make it harder to evaluate the accuracy of Lynn's hypothesis. In fact, even when g was constrained to be equal across groups (analysis not shown), the regression coefficients of age are not affected at all. Moreover, low power to detect group means is to be expected in MGCFA when

[8] In the HOF models, the verbal factor always had a g loading above .95. See the supplementals.

the group difference in latent means is small, the positive manifold is weak, and when the sample size is small (Molenaar et al., 2009). For this reason, the group equality constraint on the regression coefficient of age was ignored even though it barely affected model fit because these coefficients are not expected to differ greatly across gender groups. Generally, the reliance on statistical tests detracts from more important outcomes such as prediction accuracy and model uncertainty (Armstrong, 2007a, 2007b; Hyndman, 2008; Goodwin, 2007; Kostenko & Hyndman, 2008; Ord, 2007; Stekler, 2007). A more appropriate approach is to rely on effect sizes and confidence intervals and meta-analyses (Anderson, 2000; Hunter & Schmidt, 1996, 2004, pp. 10-11, 59-63).

An argument used to explain sex difference in g is the differential attrition affecting men, causing lower IQ men to be removed from the data during late adolescence and adulthood. As explained in the Introduction section, multiple studies could not confirm this hypothesis. In the present data, descriptive statistics (available in the supplementals) are computed for the g factor score derived from the unrotated factor analysis. Results for both the NLSY79 and NLSY97 show no change in the standard deviation (SD) of g across ages for males, whereas the attrition argument would expect a reduction of the SD at later ages for males. Moreover, a sibling pair analysis was carried out to account for selection bias resulting from SES differences between families, but it produced conflicting results. In the NLSY97, the sex difference in g and age-related g growth is consistent with the results found in the unrestricted sample. But in the NLSY79, not only the sex difference in g usually vanishes but the females showed greater age-related g growth, both of which contradict Lynn's hypothesis. While the differential attrition argument could make the point that the male advantage only vanishes in the NLSY79 due to the older age of this sample, it cannot explain why females show greater growth in g. This may be due to sampling bias.[9]

A criticism often levelled against aptitude tests is the presence of psychometric sampling bias, i.e., unbalanced sampling of cognitive dimension. In the ASVAB, tests of crystallized ability are overrepresented (Roberts et al., 2001), whereas in the Project Talent, culture-loaded knowledge tests are overrepresented (Jensen, 1985, p. 218). The observation that aptitude tests are culturally loaded and not reflecting all cognitive dimensions does not necessarily threaten their validity. Indeed, multiple studies showed that the ASVAB g factor score correlated substantially with the complexity of reaction time (RT) tests, a prototypical measure of fluid intelligence (Jensen, 1998, pp. 236-238)—and that the RT-ASVAB relationship displayed a strong correlation with ASVAB g-loadings (Larson et al., 1988) as well as with traditional IQ tests' g-loadings (Smith & Stanley, 1987). Moreover, the NLSY79 and Project Talent g factor scores displayed sizable regression coefficients for predicting educational and occupational attainment (Marks, 2023). Such a result is congruent with the meta-analytic correlation of intelligence measures with education and occupation reported by Strenze (2007).

Thus, psychometric sampling bias primarily affects the study of group means. The important lesson in this study is that both the higher order factor and bifactor models should be used for analyzing sex differences in cognitive abilities. The disappearance of group mean differences in the bifactor model suggests that sex differences in g, as observed in the higher order factor model, were actually driven by differences in specific factors. Conversely, a larger group difference in the bifactor model would indicate that g gaps were attenuated in the higher order factor model due to its dependence on non-g factors. The use of bifactor models has been questioned on the basis of its superior fit even when the true data structure is not bifactorial. However, as explained in a previous article (Hu, 2025), this pro-bifactor bias is not always present and that the bifactor is conceptually more parsimonious than the higher order factor model. Given their distinct properties, it is recommended to fit both models. This is even more important for the study of sex differences since there is still no agreement yet about the true mean differences in g. To illustrate this dilemma, the present meta-analysis of MGCFA models on IQ test batteries shows a larger male advantage in HOF models compared to BF models despite the fact that the magnitude of sex differences tend to be greater in BF models regardless of the direction of the sex differences. Studies that report estimates from both models would greatly help in solving this puzzle.

Supplementary Materials: The supplementary files can be accessed at https://osf.io/892e3/

References

Anderson, D. R., Burnham, K. P., & Thompson, W. L. (2000). Null hypothesis testing: Problems, prevalence, and an alternative. *Journal of Wildlife Management*, *64*(4), 912–923. https://doi.org/10.2307/3803199

[9] It is possible that the larger female growth is due to sample selection. Regression analyses are conducted using mother's and father's grade level as the criterion, with age, sex, and age*sex interaction as predictors. In the unrestricted sample, there is a very small, non-significant negative age*sex interaction. In the sibling sample, there is a positive age*sex interaction which shows a female advantage of about half a point in mother's grade and 1 point in father's grade when considering the full age range, despite these coefficients being non-significant. If it tells anything, it means that females tend to be more economically advantaged (i.e., more highly selected) at later ages.

Armstrong, J. S. (2007a). Significance tests harm progress in forecasting. *International Journal of Forecasting*, *23*(2), 321–327. https://doi.org/10.1016/j.ijforecast.2007.03.004

Armstrong, J. S. (2007b). Statistical significance tests are unnecessary even when properly done and properly interpreted: Reply to commentaries. *International Journal of Forecasting*, *23*(2), 335–336. https://doi.org/10.1016/j.ijforecast.2007.01.010

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. https://doi.org/10.1016/j.paid.2018.09.043

Beasley, W. H., Rodgers, J. L., Bard, D., Hunter, M. D., O'Keefe, P., Williams, K. M., & Garrison, S. M. (2024). Package 'nlsylinks'.

Beaujean, A. A. (2014). *Latent variable modeling using R: A step-by-step guide*. Routledge.

Beaujean, A. A. (2015). John carroll's views on intelligence: Bi-factor vs. higher-order models. *Journal of Intelligence*, *3*(4), 121–136. https://doi.org/10.3390/jintelligence3040121

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. https://doi.org/10.1037/a0036745

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*, *31*(1), 64–80. https://doi.org/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. https://doi.org/10.1080/10705510701301834

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. https://doi.org/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. https://doi.org/10.1207/s15328007sem0902_5

Cucina, J. M., Burtnick, S. K., Maria, E., Walmsley, P. T., & Wilson, K. J. (2024). Meta-analytic validity of cognitive ability for hands-on military job proficiency. *Intelligence*, *104*, 101818. https://doi.org/10.1016/j.intell.2024.101818

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. https://doi.org/10.1037/1082-989X.1.1.16

Deary, I. J., Irwing, P., Der, G., & Bates, T. C. (2007). Brother–sister differences in the g factor in intelligence: Analysis of full, opposite-sex siblings from the NLSY1979. *Intelligence*, *35*(5), 451–456. https://doi.org/10.1016/j.intell.2006.09.003

DeMars, C. E. (2013). A tutorial on interpreting bifactor model scores. *International Journal of Testing*, *13*(4), 354–378. https://doi.org/10.1080/15305058.2013.799067

Dolan, C. V., Colom, R., Abad, F. J., Wicherts, J. M., Hessen, D. J., & van der 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. https://doi.org/10.1016/j.intell.2005.09.003

Dykiert, D., Gale, C. R., & Deary, I. J. (2009). Are apparent sex differences in mean IQ scores created in part by sample restriction and increased male variance? *Intelligence*, *37*(1), 42–47. https://doi.org/10.1016/j.intell.2008.06.002

Eid, M. (2020). Multi-faceted constructs in abnormal psychology: Implications of the bifactor S-1 model for individual clinical assessment. *Journal of Abnormal Child Psychology*, *48*(7), 895–900. https://doi.org/10.1007/s10802-020-00624-9

Eid, M., Geiser, C., Koch, T., & Heene, M. (2017). Anomalous results in G-factor models: Explanations and alternatives. *Psychological Methods*, *22*(3), 541–562. https://doi.org/10.1037/met0000083

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*. Houghton Mifflin.

Flores-Mendoza, C., Widaman, K. F., Rindermann, H., Primi, R., Mansur-Alves, M., & Pena, C. C. (2013). Cognitive sex differences in reasoning tasks: Evidence from brazilian samples of educational settings. *Intelligence*, *41*(1), 70–84. https://doi.org/10.1016/j.intell.2012.11.002

Giofrè, D., Allen, K., Toffalini, E., & Caviola, S. (2022). The impasse on gender differences in intelligence: A meta-analysis on WISC batteries. *Educational Psychology Review*, *34*(4), 2543–2568. https://doi.org/10.1007/s10648-022-09705-1

Giofrè, D., Toffalini, E., Esposito, L., & Cornoldi, C. (2024). Sex/gender differences in general cognitive abilities: An investigation using the Leiter-3. *Cognitive Processing*, *25*(4), 663–672. https://doi.org/10.1007/s10339-024-01199-9

Goodwin, P. (2007). Should we be using significance tests in forecasting research? *International Journal of Forecasting*, *23*(2), 333–334. https://doi.org/10.1016/j.ijforecast.2007.01.008

Hambrick, D. Z., Burgoyne, A. P., & Oswald, F. L. (2024). The validity of general cognitive ability predicting job-specific performance is stable across different levels of job experience. *Journal of Applied Psychology*, *109*(3), 437–455. https://doi.org/10.1037/apl0001150

Härnqvist, K. (1997). Gender and grade differences in latent ability variables. *Scandinavian Journal of Psychology*, *38*(1), 55–62. https://doi.org/10.1111/1467-9450.00009

Harrer, M., Cuijpers, P., Furukawa, T. A., & Ebert, D. D. (2021). *Doing meta-analysis with R: A hands-on guide*. Chapman & Hall/CRC Press.

Hsu, H.-Y., Troncoso Skidmore, S., Li, Y., & Thompson, B. (2014). Forced zero cross-loading misspecifications in measurement component of structural equation models: Beware of even "small" misspecifications. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *10*(4), 138–152. https://doi.org/10.1027/1614-2241/a000084

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. https://doi.org/10.1080/10705519909540118

Hu, M. (2025). Spearman's g explains black-white but not sex differences in cognitive abilities in the project talent. *OpenPsych*. https://doi.org/10.26775/OP.2025.07.18

Hunt, E., & Madhyastha, T. (2008). Recruitment modeling: An analysis and an application to the study of male–female differences in intelligence. *Intelligence*, *36*(6), 653–663. https://doi.org/10.1016/j.intell.2008.03.002

Hunter, J. E., & Schmidt, F. L. (1996). Cumulative research knowledge and social policy formulation: The critical role of meta-analysis. *Psychology, Public Policy, and Law*, *2*(2), 324–347. https://doi.org/10.1037/1076-8971.2.2.324

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. SAGE Publications, Inc.

Hyndman, R. J. (2008). Why i don't like statistical tests. https://robjhyndman.com/hyndsight/tests/

Jensen, A. R. (1980). *Bias in mental testing*. Free Press.

Jensen, A. R. (1985). The nature of the black–white difference on various psychometric tests: Spearman's hypothesis. *Behavioral and Brain Sciences*, *8*(2), 193–219. https://doi.org/10.1017/s0140525x00020392

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

Johnson, W., Bouchard, J., T. J., Krueger, R. F., McGue, M., & Gottesman, I. I. (2004). Just one g: Consistent results from three test batteries. *Intelligence*, *32*(1), 95–107. https://doi.org/10.1016/S0160-2896(03)00062-X

Johnson, W., & Bouchard, T. J. (2007). Sex differences in mental abilities: g masks the dimensions on which they lie. *Intelligence*, *35*(1), 23–39. https://doi.org/10.1016/j.intell.2006.03.012

Johnson, W., te Nijenhuis, J., & Bouchard, J., T. J. (2008). Still just 1 g: Consistent results from five test batteries. *Intelligence*, *36*(1), 81–95. https://doi.org/10.1016/j.intell.2007.06.001

Keith, T. Z., Reynolds, M. R., Patel, P. G., & Ridley, K. P. (2008). Sex differences in latent cognitive abilities ages 6 to 59: Evidence from the Woodcock–Johnson III tests of cognitive abilities. *Intelligence*, *36*(6), 502–525. https://doi.org/10.1016/j.intell.2007.11.001

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. https://doi.org/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. https://doi.org/10.1080/10705511.2014.937791

Koch, T., & Eid, M. (2024). Augmented bifactor models and bifactor-(s-1) models are identical. a comment on zhang, luo, zhang, sun & zhang (2023). *Structural Equation Modeling: A Multidisciplinary Journal*, *31*(5), 794–801. https://doi.org/10.1080/10705511.2024.2339387

Kostenko, A. V., & Hyndman, R. J. (2008). Forecasting without significance tests.

Lakin, J., & Gambrell, J. (2014). Sex differences in fluid reasoning: Manifest and latent estimates from the cognitive abilities test. *Journal of Intelligence*, *2*(2), 36–55. https://doi.org/10.3390/jintelligence2020036

Larson, G. E., Merritt, C. R., & Williams, S. E. (1988). Information processing and intelligence: Some implications of task complexity. *Intelligence*, *12*(2), 131–147. https://doi.org/10.1016/0160-2896(88)90012-8

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. https://doi.org/10.1016/j.intell.2012.10.009

Lynn, R. (2017). Sex differences in intelligence: The developmental theory. *Mankind Quarterly*, *58*(1), 9–42. https://doi.org/10.46469/mq.2017.58.1.2

Lynn, R. (2021). *Sex differences in intelligence:the developmental theory*. Arktos Media Ltd.

Lynn, R., & Irwing, P. (2004). Sex differences on the progressive matrices: A meta-analysis. *Intelligence*, *32*(5), 481–498. https://doi.org/10.1016/j.intell.2004.06.008

Lynn, R., & Kanazawa, S. (2011). A longitudinal study of sex differences in intelligence at ages 7, 11 and 16 years. *Personality and Individual Differences*, *51*(3), 321–324. https://doi.org/10.1016/j.paid.2011.02.028

Madhyastha, T. M., Hunt, E., Deary, I. J., Gale, C. R., & Dykiert, D. (2009). Recruitment modeling applied to longitudinal studies of group differences in intelligence. *Intelligence*, *37*(4), 422–427. https://doi.org/10.1016/j.intell.2009.04.001

Marks, G. N. (2023). Has cognitive ability become more important for education and the labor market? a comparison of the project talent and 1979 national longitudinal survey of youth cohorts. *Journal of Intelligence*, *11*(8), 169. https://doi.org/10.3390/jintelligence11080169

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. https://doi.org/10.1037/0021-9010.93.3.568

Molenaar, D., Dolan, C. V., & Wicherts, J. M. (2009). The power to detect sex differences in IQ test scores using multi-group covariance and means structure analyses. *Intelligence*, *37*(4), 396–404. https://doi.org/10.1016/j.intell.2009.03.007

Moreira, J. P. A., Lopes, M. C., Miranda-Júnior, M. V., Valentini, N. C., Lage, G. M., & Albuquerque, M. R. (2019). Körperkoordinationstest für kinder (KTK) for brazilian children and adolescents: Factor analysis, invariance and factor score. *Frontiers in Psychology*, *10*, 2524. https://doi.org/10.3389/fpsyg.2019.02524

Ord, K. (2007). Comments on "significance tests harm progress in forecasting". *International Journal of Forecasting*, *23*(2), 331–332. https://doi.org/10.1016/j.ijforecast.2007.01.011

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. https://doi.org/10.1080/1091367X.2014.952370

Pesta, B. J., Bertsch, S., Poznanski, P. J., & Bommer, W. H. (2008). Sex differences on elementary cognitive tasks despite no differences on the wonderlic personnel test. *Personality and Individual Differences*, *45*(5), 429–431. https://doi.org/10.1016/j.paid.2008.05.028

Piffer, D. (2016). Sex differences in intelligence on the american WAIS-IV. *Mankind Quarterly*, *57*(1), 25–33. https://doi.org/10.46469/mq.2016.57.1.3

Pokropek, A., Davidov, E., & Schmidt, P. (2019). A monte carlo simulation study to assess the appropriateness of traditional and newer approaches to test for measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(5), 724–744. https://doi.org/10.1080/10705511.2018.1561293

Ree, M. J., & Carretta, T. R. (2022). Thirty years of research on general and specific abilities: Still not much more than g. *Intelligence*, *91*, 101617. https://doi.org/10.1016/j.intell.2021.101617

Reynolds, M. R., Hajovsky, D. B., & Caemmerer, J. M. (2022). The sexes do not differ in general intelligence, but they do in some specifics. *Intelligence*, *92*, 101651. https://doi.org/10.1016/j.intell.2022.101651

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. https://doi.org/10.1016/j.intell.2007.06.003

Roberts, R. D., Goff, G. N., Anjoul, F., Kyllonen, P. C., Pallier, G., & Stankov, L. (2000). The armed services vocational aptitude battery (ASVAB): Little more than acculturated learning (Gc)!? *Learning and individual differences*, *12*(1), 81–103. https://doi.org/10.1016/S1041-6080(00)00035-2

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

Rosén, M. (1995). Gender differences in structure, means and variances of hierarchically ordered ability dimensions. *Learning and Instruction*, *5*(1), 37–62. https://doi.org/10.1016/0959-4752(95)00002-k

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*. https://doi.org/10.31234/osf.io/wprg8

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. https://doi.org/10.3200/jexe.74.3.267-288

Smith, G. A., & Stanley, G. (1987). Comparing subtest profiles of g loadings and correlations with RT measures. *Intelligence*, *11*(4), 291–298. https://doi.org/10.1016/0160-2896(87)90012-2

Steinmayr, R., Beauducel, A., & Spinath, B. (2010). Do sex differences in a faceted model of fluid and crystallized intelligence depend on the method applied? *Intelligence*, *38*(1), 101–110. https://doi.org/10.1016/j.intell.2009.08.001

Stekler, H. (2007). Significance tests harm progress in forecasting: Comment. *International Journal of Forecasting*, *23*(2), 329–330. https://doi.org/10.1016/j.ijforecast.2007.01.009

Strenze, T. (2007). Intelligence and socioeconomic success: A meta-analytic review of longitudinal research. *Intelligence*, *35*, 401–426. https://doi.org/10.1016/j.intell.2006.09.004

van der Linden, W. J., & Glas, C. A. (Eds.). (2000). *Computerized adaptive testing: Theory and practice*. Kluwer Academic.

van der Sluis, S., Derom, C., Thiery, E., Bartels, M., Polderman, T. J., Verhulst, F. C., Posthuma, D., et al. (2008). Sex differences on the WISC-R in belgium and the netherlands. *Intelligence*, *36*(1), 48–67. https://doi.org/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. https://doi.org/10.1016/j.intell.2005.08.002

Watts, A. L., Poore, H. E., & Waldman, I. D. (2019). Riskier tests of the validity of the bifactor model of psychopathology. *Clinical Psychological Science*, *7*(6), 1285–1303. https://doi.org/10.1177/2167702619855035

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. https://doi.org/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. https://doi.org/10.3389/fpsyg.2022.923877

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. https://doi.org/10.1080/00273171.2021.1957664

Zhou, L. (2023). Distinguish the bifactor and higher-order factor model: A comparison of three RMSEA-related approaches under model misspecification (T). https://open.library.ubc.ca/collections/ubctheses/24/items/1.0437126

Appendix

Table A1 displays the results for non-Hispanic Whites in the NLSY79. In both the CF and HOF models, metric invariance was tenable but scalar invariance was seriously violated. A partial scalar model showed an acceptable change in fit only after freeing 4 intercepts (WK, MK or AR, EI, CS) in CF and HOF models and 1 intercept (PC) in the BF model. In the case of the HOF model, MK was freed instead of AR subtest because it led to a final model that was slightly more parsimonious, without issues related to the non-positive definiteness of the variance-covariance matrix, with almost no difference in model fit.[a] In the CF model, the group equality of latent covariances caused a deterioration in fit due to a large change in SRMR. This constraint is however necessary because freeing the latent covariances implies that there are group differences in the construct. The next step involves the equality of factor variances. But only a partial invariance could achieve acceptable fit, for all 3 models. In this analysis, the CF model fits better than both the HOF and BF models based on CFI, RMSEA and Mc, whereas the HOF and BF models fit better than CF based on SRMR. On one hand, it suggests that neither *g* model is the best representation of the data. On the other hand, in the context of MGCFA, it is not clear that the CF model should be the best model due to lack of invariance in the latent covariances. A group difference in shared variability across the constructs may reflect group differences in cognitive strategies or developmental processes, which would threaten group comparability in latent scores.

Table A1: MGCFA models for gender groups in the NLSY79 (White sample)

Model Level	χ^2	df	CFI	RMSEA	SRMR	Mc	RMSEA_D [CI]
Correlated Factors Model							
M0. Baseline	1805	27	.961	.103	.040	.865	
M1. Configural	1104	54	.976	.079	.024	.918	
M2. Metric	1250	62	.973	.079	.031	.908	.062 [.052 : .073]
M3. Scalar	2462	68	.946	.107	.063	.823	.210 [.198 : .222]
M3a. Partial Scalar[1]	1291	64	.972	.079	.033	.905	.063 [.043 : .085]
M4. Lv covariance	1408	70	.970	.079	.093	.897	.063 [.051 : .076]
M5. Lv var-covariance[2]	1417	73	.970	.077	.092	.897	.020 [NA : .041]
M6. Lv reduced[3]	**1425**	**74**	**.970**	**.077**	**.092**	**.896**	**.037 [.012 : .071]**
Higher Order Factor Model							
M0. Baseline	2507	29	.946	.118	.058	.818	
M1. Configural	1639	58	.964	.094	.034	.879	
M2. Metric	1806	69	.961	.090	.044	.868	.056 [.047 : .065]

Model Level	χ^2	df	CFI	RMSEA	SRMR	Mc	$RMSEA_D$ [CI]
M3. Scalar	2989	74	.934	.113	.071	.789	.210 [.197 : .223]
M3a. Partial Scalar[1]	1874	70	.959	.091	.046	.864	.086 [.058 : .117]
M4. Lv variance[2]	1881	73	.959	.090	.047	.863	.014 [NA : .036]
M5. Lv reduced[3]	**1881**	**74**	**.959**	**.089**	**.047**	**.863**	**NaN***
Bifactor Model							
M0. Baseline	1976	26	.957	.110	.048	.853	
M1. Configural	1460	52	.968	.094	.033	.892	
M2. Metric	1560	67	.966	.085	.041	.886	.036 [.028 : .044]
M3. Scalar	1846	73	.960	.089	.043	.866	.113 [.101 : .126]
M3a. Partial Scalar[1]	1675	72	.964	.085	.042	.878	.080 [.067 : .094]
M4. Lv variance[2]	**1677**	**74**	**.964**	**.084**	**.042**	**.878**	**NaN***

[1] Freed parameters: intercept of WK, AR, EI, CS for CF model; intercept of WK, MK, EI, CS for HOF model; intercept of PC for BF model.
[2] Freed parameters: variance of the verbal, math, and speed factors.
[3] Fixed parameters: intercept of the math factor.
* NaN is the result of a chi-square that is negative or lower than 1 (model fits better). In this case, $RMSEA_D$ cannot be computed.
[a] The exception is for $RMSEA_D$, which produced values of .086 and .066 for free MK and free AR, respectively, for partial scalar models.

Table A2 displays the results for the sample comprising all races in the NLSY79. Metric invariance was tenable for CF but not HOF model. A partial metric specification for the HOF model achieved good fit after freeing the loading of the electronic factor on *g*. Scalar invariance was not tenable for all 3 competing models, and the change in fit was acceptable only after freeing 3 intercepts for CF and HOF models and 1 intercept for the BF model. In the CF model, the group equality of latent covariances caused a huge misfit in the SRMR. The next step involves the equality of factor variances. But only a partial invariance could achieve acceptable fit, for all 3 models. In this analysis, the CF model fits better than the HOF based on CFI, RMSEA and Mc, but the advantage is quite modest, whereas the HOF model fits better than CF based on SRMR. The BF model fits just as well as the CF model except for SRMR where it displays a better fit.

Table A2: MGCFA models for gender groups in the NLSY79 (all race sample)

Model Level	χ^2	df	CFI	RMSEA	SRMR	Mc	$RMSEA_D$ [CI]
Correlated Factors Model							
M0. Baseline	2636	27	.973	.094	.029	.887	
M1. Configural	1759	54	.982	.076	.019	.925	
M2. Metric	1983	62	.980	.075	.026	.916	.056 [.049 : .065]
M3. Scalar	3572	68	.963	.097	.046	.852	.174 [.165 : .183]
M3a. Partial Scalar[1]	2235	65	.977	.078	.029	.905	.096 [.083 : .109]
M4. Lv covariance	2625	71	.973	.081	.108	.890	.089 [.080 : .098]
M5. Lv var-covariance[2]	**2674**	**74**	**.972**	**.080**	**.107**	**.888**	**.042 [.029 : .055]**
Higher Order Factor Model							
M0. Baseline	3823	29	.960	.109	.044	.840	
M1. Configural	2545	58	.974	.089	.026	.892	
M2. Metric	2992	69	.969	.088	.052	.875	.070 [.064 : .077]
M2a. Partial Metric[3]	2744	68	.972	.085	.032	.884	.048 [.041 : .055]
M3. Scalar	4294	73	.955	.103	.051	.824	.173 [.163 : .183]
M3a. Partial Scalar[1]	2992	70	.969	.087	.034	.875	.096 [.081 : .113]
M4. Lv variance[2]	3002	73	.969	.086	.034	.874	.015 [NA : .030]
M5. Lv reduced[4]	**3002**	**74**	**.969**	**.085**	**.034**	**.874**	**NA**
Bifactor Model							
M0. Baseline	3084	26	.968	.104	.037	.869	

Model Level	χ^2	df	CFI	RMSEA	SRMR	Mc	$RMSEA_D$ [CI]
M1. Configural	2280	52	.976	.089	.026	.903	
M2. Metric	2635	67	.973	.084	.053	.889	.053 [.047 : .059]
M3. Scalar	3075	73	.968	.087	.054	.871	.101 [.092 : .110]
M3a. Partial Scalar[1]	2817	72	.971	.084	.054	.882	.072 [.062 : .083]
M4. Lv variance[2]	**2824**	**74**	**.971**	**.083**	**.054**	**.882**	**.015 [NA : .034]**

[1] Freed parameters: intercept of WK, EI, CS for CF and HOF models; intercept of PC for BF model.
[2] Freed parameters: variance of the verbal, math, and speed factors for CF and HOF models; variance of math and speed factors for BF model.
[3] Freed parameters: loading of electronic on *g*.
[4] Fixed parameters: intercept of the math factor.

Table A3 displays the results for non-Hispanic Whites in the NLSY97. Metric invariance was tenable for CF but barely tenable for HOF and BF models. The major source of non-invariance was the loading of EI on the electronic factor for the HOF model and the loading of EI on *g* for the BF model. Scalar invariance was not tenable for all 3 competing models, and the change in fit was acceptable only after freeing 3 intercepts for CF and HOF models and 4 intercepts for BF model. In the CF model, the group equality of latent covariances caused a huge misfit in the SRMR. The next step involves the equality of factor variances. But only a partial invariance could achieve acceptable fit, for all 3 models. In this analysis, the CF model fits much better than the HOF model based on CFI, RMSEA and Mc, whereas the HOF model fits better than the CF model based on SRMR. The BF model fits just as well as the CF model except for SRMR where it displays a better fit.

Table A3: MGCFA models for gender groups in the NLSY97 (White sample)

Model Level	χ^2	df	CFI	RMSEA	SRMR	Mc	$RMSEA_D$ [CI]
Correlated Factors Model							
M0. Baseline	1095	45	.967	.080	.030	.866	
M1. Configural	917	90	.974	.071	.026	.893	
M2. Metric	1026	101	.971	.071	.038	.881	.062 [.051 : .075]
M3. Scalar	1611	109	.953	.087	.044	.814	.191 [.177 : .205]
M3a. Partial Scalar[1]	1120	106	.968	.072	.040	.870	.092 [.075 : .110]
M4. Lv covariance	1224	112	.965	.074	.095	.859	.088 [.072 : .104]
M5. Lv var-covariance[2]	1228	115	.965	.073	.095	.859	NA
M6. Lv reduced	**1229**	**117**	**.965**	**.072**	**.095**	**.859**	**NA**
Higher Order Factor Model							
M0. Baseline	1514	47	.954	.092	.044	.818	
M1. Configural	1241	94	.964	.082	.036	.855	
M2. Metric	1373	108	.960	.080	.050	.841	.060 [.050 : .071]
M2a. Partial Metric[3]	1304	107	.962	.078	.041	.849	.040 [.029 : .052]
M3. Scalar	1879	114	.944	.092	.047	.786	.188 [.174 : .203]
M3a. Partial Scalar[1]	1390	111	.960	.079	.043	.839	.087 [.069 : .108]
M4. Lv variance[2]	**1400**	**113**	**.959**	**.079**	**.043**	**.839**	**.035 [.008 : .067]**
Bifactor Model							
M0. Baseline	1302	43	.961	.089	.045	.842	
M1. Configural	1073	86	.969	.079	.036	.874	
M2. Metric	1207	105	.965	.076	.051	.860	.051 [.042 : .061]
M2a. Partial Metric[3]	1125	104	.968	.073	.041	.870	.027 [.017 : .038]
M3. Scalar	1679	112	.951	.087	.046	.807	.351 [.337 : .365]
M3a. Partial Scalar[1]	1162	108	.967	.073	.042	.866	.064 [.046 : .085]
M4. Lv variance[2]	1171	110	.967	.073	.042	.865	.031 [NA : .064]
M5. Lv reduced[4]	**1175**	**111**	**.966**	**.072**	**.043**	**.865**	**.037 [NA : .083]**

[1] Freed parameters: intercept of PC, NO, GS for CF and HOF models; intercept of AR, GS, CS, PC for BF model.
[2] Freed parameters: variance of the verbal, math, and speed factors for CF model; variance of speed and math factors for HOF model; variance of verbal and speed factors for BF model.
[3] Freed parameters: loading of EI on electronic factor for HOF model; loading of EI on *g* for BF model.
[4] Freed parameters: intercept of speed factor for BF model.

Table A4 displays the results for the sample comprising all races in the NLSY97. Metric invariance was tenable for CF and BF models but barely tenable for the HOF model. A partial metric specification for the HOF model achieved good fit after freeing the loading of the EI subtest on the electronic factor. Scalar invariance was not tenable for all 3 competing models, and the change in fit was acceptable only after freeing 2 intercepts for CF and HOF models and 3 intercepts for the BF model. In the CF model, the group equality of latent covariances caused a huge misfit in the SRMR. The next step involves the equality of factor variances. But only a partial invariance could achieve acceptable fit, for all 3 models. In this analysis, overall, the CF model fits much better than the HOF model based on CFI, RMSEA and Mc, whereas the HOF model fits better than the CF model based on SRMR. The BF model fits just as well as the CF model except for SRMR where it displays a better fit.

Table A4: MGCFA models for gender groups in the NLSY97 (all race sample)

Model Level	χ^2	df	CFI	RMSEA	SRMR	Mc	RMSEA_D [CI]
Correlated Factors Model							
M0. Baseline	2017	45	.971	.079	.027	.870	
M1. Configural	1738	90	.976	.072	.025	.890	
M2. Metric	1915	101	.973	.071	.034	.880	.056 [.048 : .065]
M3. Scalar	2799	109	.960	.083	.039	.827	.158 [.149 : .168]
M3a. Partial Scalar[1]	2132	107	.970	.073	.036	.867	.088 [.077 : .100]
M4. Lv covariance	2367	113	.967	.075	.094	.853	.091 [.080 : .103]
M5. Lv var-covariance[3]	2371	116	.967	.074	.094	.853	NA
M6. Lv reduced[4]	**2371**	**117**	**.967**	**.074**	**.094**	**.853**	**NA**
Higher Order Factor Model							
M0. Baseline	2853	47	.959	.092	.040	.820	
M1. Configural	2415	94	.966	.083	.033	.849	
M2. Metric	2673	108	.962	.082	.050	.834	.060 [.053 : .068]
M2a. Partial Metric[2]	2538	107	.964	.080	.042	.842	.041 [.034 : .050]
M3. Scalar	3410	114	.951	.090	.046	.793	.156 [.146 : .167]
M3a. Partial Scalar[1]	2753	112	.961	.082	.043	.830	.088 [.076 : .101]
M4. Lv variance[3]	2762	114	.961	.081	.043	.830	.024 [NA : .047]
M5. Lv reduced[4]	**2762**	**115**	**.961**	**.081**	**.043**	**.830**	**NA**
Bifactor Model							
M0. Baseline	2403	42	.965	.089	.040	.847	
M1. Configural	2015	84	.971	.081	.032	.873	
M2. Metric	2296	104	.968	.077	.049	.857	.051 [.044 : .057]
M3. Scalar	3136	112	.955	.087	.053	.808	.159 [.149 : .169]
M3a. Partial Scalar[1]	2369	109	.967	.076	.050	.853	.054 [.042 : .067]
M4. Lv variance[3]	**2369**	**110**	**.967**	**.076**	**.050**	**.853**	**NA**

[1] Freed parameters: intercept of PC and NO for CF and HOF model; intercept of PC, NO and WK for BF model.
[2] Freed parameters: loading of EI on the electronic factor for HOF model.
[3] Freed parameters: variance of the verbal, math, and speed factors for CF model; variance of math and speed factors for HOF model; variance of the math factor for BF model.
[4] Fixed parameters: intercept of the math factor for CF model; intercept of the verbal factor for HOF model.

Table A5 displays the results for the White sample in the Project Talent. Metric invariance was tenable for CF and HOF models but not for the BF model. A partial metric specification for the BF model achieved good fit

after freeing the loading of the vocabulary on *g*. Scalar invariance was not tenable for all 3 competing models, and the change in fit was acceptable only after freeing 11 intercepts for CF and HOF models and 8 intercepts for the BF model. In the CF model, the equality constraint was imposed on latent covariances despite a huge misfit observed in the SRMR value. The next step involves testing the equality of factor variances. This led to a small misfit in Mc for the HOF and BF models, but no misfit was observed in the CF model. In this analysis, the CF model fits better than the HOF model but worse than the BF model.

Table A5: MGCFA models for gender groups in the Project Talent (White sample)

Model Level	χ^2	df	CFI	RMSEA	SRMR	Mc	RMSEA_D [CI]
Correlated Factors Model							
M0. Baseline	165505	495	.941	.048	.041	.560	
M1. Configural	127095	990	.953	.042	.032	.642	
M2. Metric	135951	1035	.950	.043	.040	.622	.049 [.048 : .050]
M3. Scalar	203044	1063	.925	.052	.046	.491	.170 [.169 : .171]
M3a. Partial Scalar[1]	144975	1052	.947	.044	.041	.603	.080 [.078 : .081]
M4. Lv covariance	148199	1067	.945	.044	.058	.596	.052 [.050 : .053]
M5. Lv var-covariance	149594	1073	.945	.044	.059	.593	.051 [.049 : .054]
M6. Lv reduced[3]	**149712**	**1074**	**.945**	**.044**	**.059**	**.593**	**.038 [.032 : .044]**
Higher Order Factor Model							
M0. Baseline	219645	504	.922	.055	.057	.463	
M1. Configural	145152	1008	.946	.045	.037	.602	
M2. Metric	154823	1058	.943	.045	.045	.582	.049 [.048 : .050]
M3. Scalar	223725	1085	.917	.054	.050	.457	.172 [.171 : .173]
M3a. Partial Scalar[1]	165246	1074	.939	.046	.046	.561	.086 [.084 : .088]
M4. Lv variance	168139	1081	.938	.047	.062	.556	.071 [.068 : .073]
M5. Lv reduced[3]	**168139**	**1082**	**.938**	**.047**	**.062**	**.556**	**.002 [NA : .011]**
Bifactor Model							
M0. Baseline	173063	476	.939	.051	.052	.545	
M1. Configural	113410	952	.958	.041	.028	.673	
M2. Metric	124909	1030	.954	.041	.038	.647	.042 [.041 : .042]
M2a. Partial Metric[2]	122819	1029	.955	.041	.036	.651	.038 [.037 : .039]
M3. Scalar	153193	1056	.943	.045	.040	.585	.122 [.120 : .123]
M3a. Partial Scalar[1]	129910	1048	.952	.042	.038	.635	.064 [.063 : .066]
M4. Lv variance	**132591**	**1055**	**.951**	**.042**	**.058**	**.630**	**.067 [.065 : .070]**

[1] Freed parameters: intercept of social studies, theater, law, music, physical science, miscellaneous, visual3D, electronics, math, health, biological science for CF model; intercept of social studies, theater, law, music, miscellaneous, health, physical science, electronics, visual3D, biological science, vocabulary for HOF model; intercept of disguised words, physical science, health, electronics, visual3D, vocabulary, art, biological science for BF model. By decreasing order of effect.
[2] Freed parameters: loading of vocabulary on *g*.
[3] Fixed parameters: intercept of the info factor for CF model; intercept of the math factor for HOF model.

Table A6 displays the results for the Black sample in the Project Talent. Metric invariance was untenable for both CF and BF models, and a partial metric specification for the BF model achieved good fit after freeing the most offending parameters. Scalar invariance was not tenable for both CF and BF models, and the change in fit was acceptable only after freeing 5 intercepts and 4 intercepts for the CF and BF models respectively. In the CF model, the equality constraint was once again imposed on latent covariances despite a noticeable misfit in all fit indices (in particular, RMSEA_D). The next step involves testing the equality of factor variances. This led to a small misfit in Mc for the HOF and BF models, but no misfit was observed in the CF model.

Table A6: MGCFA models for gender groups in the Project Talent (Black sample)

Model Level	χ^2	df	CFI	RMSEA	SRMR	Mc	$RMSEA_D$ [CI]
Correlated Factors Model							
M0. Baseline	6177	500	.950	.043	.039	.627	
M1. Configural	5468	1000	.960	.038	.035	.693	
M2. Metric	6284	1040	.953	.041	.052	.650	.075 [.070 : .080]
M2a. Partial Metric[1]	5862	1037	.957	.039	.044	.673	.053 [.048 : .058]
M3. Scalar	7847	1065	.939	.046	.056	.573	.146 [.140 : .152]
M3a. Partial Scalar[2]	6407	1060	.952	.041	.045	.644	.081 [.075 : .088]
M4. Lv covariance	6919	1075	.948	.042	.079	.619	.095 [.088 : .103]
M5. Lv var-covariance	7028	1081	.947	.043	.081	.613	.065 [.053 : .078]
M6. Lv reduced[3]	**7028**	**1082**	**.947**	**.043**	**.081**	**.613**	**NA**
Higher Order Factor Model							
M0. Baseline	7272	509	.940	.047	.044	.574	
M1. Configural	6061	1018	.955	.040	.038	.661	
M2. Metric	7254	1063	.944	.044	.066	.601	.086 [.082 : .091]
M2a. Partial Metric[1]	6495	1059	.951	.041	.048	.640	.053 [.048 : .057]
M3. Scalar	8481	1086	.934	.047	.061	.544	.146 [.140 : .152]
M3a. Partial Scalar[2]	7028	1081	.947	.043	.049	.613	.080 [.074 : .087]
M4. Lv variance	7091	1088	.946	.043	.057	.611	.044 [.033 : .056]
M5. Lv reduced[3]	**7091**	**1089**	**.946**	**.043**	**.057**	**.611**	**NA**
Bifactor Model							
M0. Baseline	5831	481	.953	.043	.034	.644	
M1. Configural	4787	962	.966	.036	.025	.730	
M2. Metric	5968	1035	.956	.040	.061	.667	.064 [.060 : .067]
M2a. Partial Metric[1]	5341	1029	.961	.037	.041	.702	.044 [.040 : .048]
M3. Scalar	6509	1056	.951	.041	.046	.639	.110 [.104 : .116]
M3a. Partial Scalar[2]	5802	1052	.957	.039	.042	.677	.072 [.066 : .078]
M4. Lv variance	**5976**	**1059**	**.956**	**.039**	**.058**	**.667**	**.079 [.068 : .091]**

[1] Freed parameters: loading of aeronautics and social studies on info, mechanical reasoning on spatial for CF model; loading of science on *g*, aeronautics and social studies on info, mechanical reasoning on spatial for HOF model; loading of aeronautics, mechanical reasoning, mechanics, electronics, physical science, social studies on *g* for BF model.

[2] Freed parameters: intercept of mechanics, aeronautics, mechanical reasoning, social studies, theater for both the CF and HOF models; intercept of mechanical reasoning, aeronautics, disguised words, health for BF model. By decreasing order of effect.

[3] Fixed parameters: intercept of the info factor for CF model; intercept of the math factor for HOF model.