psych

Supplementary File

Section 1: Psychometric Assessment of the Penn Computerized Neurocognitive Battery

1.1. Spearman's Hypothesis and the Method of Correlated Vectors

Spearman's hypothesis (SH), as formulated by Jensen [1], is the proposition that the AA-EA (black-white) group differences on tests of cognitive ability are primarily due to the *g* factor. SH can be thought of as taking two different forms: strong and weak. According to the strong form, the observed group differences are entirely a result of differences in *g* and not related to other, specific abilities. According to the weak form, endorsed by Jensen [1–3], group differences are only primarily due to differences in *g*. Both of these manifestations of SH can be contrasted with an alternative contra form, according to which the group differences "resides entirely or mainly in the tests' group factors and specificity" and to which the "*g* factor contributes little or nothing" ([2], p. 372).

Spearman's hypothesis is typically assessed using a technique called the Method of Correlated Vectors (MCV). This method involves correlating the vector of subtest or item *g* loadings of an assessment battery with the vector of group differences for those items or subtests. The finding of a sizeable positive relationship between the vector of *g* loadings and the vector of group difference is taken to indicate a "Jensen effect," whereas a negative or null relationship indicates that something is either an anti-Jensen effect or a non-Jensen effect [4,5]. Jensen effects are taken as evidence for Spearman's hypothesis [6]. However, the MCV has been criticized as a tool for evaluating Spearman's hypothesis because it lacks specificity and sensitivity [7,8], calling into question the validity of this interpretation of Jensen effects.

1.2. An Alternative Method

Some scholars have proposed using a more advanced technique called multi-group confirmatory factor analysis (MGCFA) [7,9–11] to assess the causes of group differences. MGCFA has the advantage of allowing scholars to assess measurement invariance (MI) where MCV-based analyses, at best, use congruence coefficients to assess bias. Mellenbergh [12] defined a test as unbiased when the following condition was met:

$$f(\mathrm{Y} \mid \eta) = f(\mathrm{Y} \mid \eta, \mathrm{s})$$

where Y and η are observed and factor scores, respectively while s is group membership. Thus, given someone's latent score, η, the observed score, Y, does not depend on group membership. When observed scores are solely a function of factor scores and are independent of group membership, MI is said to hold [11,13–15]. Strict factorial invariance (SFI), where the residuals are constrained to equality in both groups, implies that the group differences are a subset of the within-group differences [11,15]. MI is typically assessed by fitting identical models of the assessment in question to different groups and then constraining these models in predefined steps [16]. These steps are described in Table S1 below.

Table S1. Steps in Assessing Measurement Invariance.

Model	Name	Constraint	Free Parameters	Implication
1	Configural, Pattern, or Form Invariance	Latent means and variances	The rest	Same number of latent variables, indicators, and pattern of constrained and estimated parameters

2	Metric or Weak Invariance	Latent means, factor loadings	Intercepts, latent variances	Same underlying construct meanings, comparable latent variances and covariances
3	Scalar or Strong Invariance	Intercepts	Latent means	Comparable latent means
4	Strict (Factorial), Omnibus, or Full Uniqueness Invariance	Error variances	-	Identical latent variables and reliability
5	Homogeneity of Latent Variances	Latent variances	-	Groups use equivalent ranges of the latent construct
6	Homogeneity of Factor Means	Latent Means	-	No difference between groups in the level of the latent construct

Note: Constraints are added to the prior level and thus models are "nested".

Model fit is assessed with multiple indices including the comparative fit index (CFI), root mean square error of approximation (RMSEA), McDonald's noncentrality index (Mc), the Tucker-Lewis Index (TLI), and the standardized root mean square residual (SRMR). χ2/df is one of the most commonly-used measures but we note that this statistic penalizes large sample sizes [17,18] and our sample sizes are large.

1.3. MGCFA of the Philadelphia Computerized Neurocognitive Battery

It is our aim to test for bias and assess Spearman's hypothesis in the Philadelphia Computerized Neurocognitive Battery (PCNB) [19]. Hu & Bentler's [20] adequate model fit measures (CFI ≥ 0.95, Mc ≥ 0.9, and RMSEA ≤ 0.06) were used to find an acceptable initial model. Our criteria for determining non-invariance come from Cheung and Rensvold [21] and Chen [22]. The former argued that a ΔCFI of greater than −0.010 or a ΔMc greater than -0.020 between nested models indicates violation of measurement invariance while the latter showed using simulations that a change of ≥-0.010 in CFI coupled with a change of ≥0.015 in RMSEA is a good criteria for large and equal samples, while a change of ≥ −0.005 in CFI coupled with ≥0.010 in RMSEA is good for small and unequal sample sizes. (cf. [23,24]). We regard a ΔCFI of greater than −0.01, a ΔMc greater than −0.02, and a ΔRMSEA greater than 0.01 as evidence that measurement invariance is untenable.

A critical assumption underlying MGCFA is that observed variables are multivariate normally distributed [25]. Univariate skewness values were < 2 for all tests except for the LNB and all kurtosis values were < 7 except for the LNB, and so were deemed acceptable [26]. The LNB values were only barely non-normal (skewness −2.21, kurtosis 8.31) and log transformation put them within the normal range (−1 and 2.43) without substantially reducing the correlation to *g*. Multivariate skew and kurtosis was minor (b1p = 14.71; b2p = 127.81) and a Q-Q plot revealed that the distribution differed little from a normal one. Winsorization (Dixon, 1980) [27] or removal of responses deemed to be outliers (n = 82, 37 black and 45 white, or 1% of the sample) eliminated any violation of non-normality. Our data thus approximate a multivariate normal distribution and it is appropriate to conduct an MGCFA. We settled on a bifactor model [10,28]. Our initial MGCFA results using the theoretical model from Moore et al. [19] minus the social cognition factor are presented in Table S2 below.

Table S2. Theory Bifactor Solution for African and European Americans on the Philadelphia Computerized Neurocognitive Battery.

Model	MI Step	χ2 /df	CFI	ΔCFI	RMSEA	ΔRMSEA	Mc	ΔMc	SRMR
0	Baseline	6.290	0.972	-	0.025	-	0.979	-	0.019
1	Configural	4.577	0.984	0.012	0.021	−0.004	0.988	0.009	0.016
2	Metric	4.556	0.979	−0.005	0.021	0	0.985	−0.003	0.018
3	Scalar	5.975	0.969	−0.010	0.025	0.004	0.977	−0.008	0.022

4	Strict	6.866	0.959	−0.010	0.031	0.006	0.970	−0.007	0.024

Note: Combined *N* = 8,143, with 3,067 African-Americans and 5,076 European-Americans Baseline is Three Stratum-II Factors without *g*.

It is worth noting that the addition of a *g* factor considerably improved model fit relative to the model without it. When *g* was added, the factor loadings for Executive Functioning, Complex Cognition, and Episodic Memory decreased; additionally, subtest *g* loadings were virtually identical regardless of the subtests included in the analysis, as assessed by comparing models after adding and removing each subtest. Deterioration in model fit at the configural and metric stage was acceptable. At the scalar and strict phase, ΔCFI was −0.10. According to Cheung & Rensvold [21] this is acceptable because it is not greater than −0.10. RMSEA never increased by more than 0.006 and the Mc never decreased by more than 0.008, while the CFI exactly touched our threshold of -0.010 in the Scalar and Strict stages.

Nonetheless, we reassessed the models with partial invariance, additional residual covariances, and using the variables after winsorization [27]. The winsorized data had a maximum ΔCFI of −0.008 (for both the scalar and strict steps) and a model with covarying residual errors for the PLOT and PVRT and PLOT and PMAT tests had a deterioration in fit of ΔCFI -0.006 (CFI = 0.973) at the scalar step and −0.008 (CFI = 0.965) at the strict step. Further, freeing the PWMT (all subtest variables) residual in the strict step led to a ΔCFI of −0.007 (CFI = 0.966). Using the Winsorized data in conjunction with the model with covaried errors and the freed PWMT residual led to even smaller deterioration in fit (maximum ΔCFI = −0.006). Using the R package *blavaan* [29] to fit an approximate measurement invariance model [33], the scalar ΔCFI becomes -0.004. MI, as SFI, was deemed tenable.

To be sure of this conclusion, an EFA was conducted resulting in a four factor model with a factor for the PCET, PLOT, PMAT, PVRT, and VOLT, a factor for the WRAT and PVRT, a factor for the VOLT, PWMT, and PFMT, and a factor for the PCPT, LNB, PVRT, and PCET, and a second round of MGCFA was done based on this. This had substantially better fit; the results can be seen in Table S3.

Table S3. EFA Bifactor Solution for African and European Americans on the Philadelphia Computerized Neurocognitive Battery.

Model	MI Step	χ2 /df	CFI	ΔCFI	RMSEA	ΔRMSEA	Mc	ΔMc	SRMR
0	Baseline	5.516	0.981	-	0.027	-	0.986	-	0.021
1	Configural	5.345	0.989	0.008	0.023	−0.004	0.992	0.006	0.015
2	Metric	4.143	0.986	-0.003	0.020	−0.003	0.989	−0.003	0.020
3	Scalar	4.099	0.985	-0.001	0.019	−0.001	0.989	0	0.020
4	Strict	4.917	0.980	-0.005	0.029	0.010	0.971	−0.018	0.033

Note: Combined N = 8,143, with 3,067 African-Americans and 5,076 European-Americans. Baseline is three stratum-II factors without g.

Spearman's hypothesis was assessed with latent variances fixed (which improved model fit) in both groups in the theory-based model. Both the strong (only group factors constrained to zero) and contra (as either only g constrained to zero or as g and episodic memory) hypotheses were rejected in terms of ΔCFI and ΔMc, as the former statistic was always > −0.01 and the latter was always > −0.02. The weak model was assessed by constraining only episodic memory, which is known to favor blacks net of *g* [1,2]. This model barely deteriorated (ΔCFI = −0.003; ΔMc = −0.004), but this is perhaps to be expected because this is effectively a much smaller constraint which is fitting for a theoretically much more relaxed model. As a result of this finding, the black-white cognitive ability difference in our data can be thought of as a product of differences in both *g* and specific abilities. Differences in the various abilities in the homogeneous latent variances model without additional constraints are given in Table S4 and differences in the weak Spearman's model are given in Table 1 in the main paper.

Table S4. Factor Score Differences between African- and European-Americans based on the Model with Constrained Latent Variances.

Factor	Estimate	SE	Lower 95% CI	Upper 95% CI
g	1.108	0.020	1.069	1.147
Complex Cognition	0.322	0.027	0.269	0.375
Executive Functioning	−0.095	0.029	−0.152	−0.038
Episodic Memory	−0.560	0.020	−0.599	−0.521

Note: Positive values indicate higher white scores and vice-versa. Estimates are in terms of Cohen's d.

Our findings accord with the literature referenced in the main paper in that we find a black-white difference of roughly 1.1 d. Furthermore, they illustrate that the residual gap (that is, without *g*) is small, favors blacks for some abilities, and does so especially strongly for our Episodic Memory factor, consistent with earlier findings. Given that *g* is central to group differences in our study, we regard Jensen effects as interpretable in the normal sense they're used in the literature. See [31] for more information on g-based inferences. Final statistical notes for this section of the study include that EFA factor loadings had congruence coefficients around 0.99 between groups and correlated at that level with the MGCFA loadings, the ωh for this battery was 0.69, reaction time computed from the 50 RT measures in the study correlated at 0.21 with *g*, the Scarr-Rowe effect defined as the difference in heritabilities between blacks and whites (taken from Mollon et al., [32]) correlated at -0.35 with subtest *g* loadings which indicates an anti-Jensen effect, and the group differences from an EFA correlated at 0.99 with the group differences derived from MGCFA.

1.4. MGCFA of the NIH Toolbox Cognition Battery

Kirkegaard et al. [33] evaluated the relationship between admixture and cognitive ability differences in the nationally-representative Pediatric Imaging, Neurocognition, and Genetics Study (PING; n = 1369). Their sample with genetic and test battery information included individuals who self-identified as White (n = 567), African-American (138), American Indian (4), Asian (120), Hispanic (323), Multi-ethnic (182), Other (19), and Pacific Islander (16), with an average age of 11.75 years. These authors did not assess whether their battery was MI, though they did report congruence coefficients which were greater than 0.9 for all group comparisons except Asians and AAs. We assess and report MI for their sample using data from AAs, EAs, and HAs noting that all univariate skewness and kurtosis coefficients are below 3 and the b1p is 10.04 and the b2p is 81.59 for their combined group. The model is derived from an EFA of the seven tests used in their study which yields three factors, verbal, spatial, and memory. After assessing MI, Spearman's hypothesis is tested separately for the AA-EA and HA-EA models. These results are given in Tables S6 and S7.

Table S6. Bifactor Solution for African and non-Hispanic White Americans on the Pediatric Imaging, Neurocognition, and Genetics Study Battery.

Model	MI Step	$\chi 2$ /df	CFI	ΔCFI	RMSEA	ΔRMSEA	Mc	ΔMc	SRMR
1	Configural	3.27	0.996	-	0.060	-	0.989	-	0.010
2	Metric	1.98	0.995	−0.001	0.040	−0.020	0.984	−0.005	0.037
3	Scalar	1.74	0.995	0	0.034	−0.006	0.986	0.002	0.033
4	Strict	1.94	0.992	−0.003	0.039	0.005	0.978	−0.008	0.044

Note: Combined n = 1140, with 228 African-Americans and 912 European-Americans.

Table S7. Bifactor Solution for Hispanic and non-Hispanic White Americans on the Pediatric Imaging, Neurocognition, and Genetics Study Battery.

Model	MI Step	χ2 /df	CFI	ΔCFI	RMSEA	ΔRMSEA	Mc	ΔMc	SRMR
1	Configural	3.176	0.997	-	0.053	-	0.992	-	0.009
2	Metric	1.316	0.999	0.002	0.020	−0.033	0.996	0.004	0.026
3	Scalar	1.304	0.999	0	0.020	0	0.995	−0.001	0.024
4	Strict	2.65	0.991	−0.008	0.042	0.022	0.971	−0.016	0.035

Note: Combined n = 1240, with 328 Hispanic-Americans and 912 European-Americans.

These models are unbiased and they fit well. The NIH Toolbox Cognition Battery can be considered an unbiased assessment for native English-speaking Hispanic and African Americans (see Wicherts & Dolan [34] on the effects of language bias). This study should serve as another datapoint in a future meta-analysis of group differences in latent cognitive ability. To note, the black-white difference was 0.97 d and the Hispanic-white difference was 0.62 d. SES in this sample was slightly above the national average.

Regarding the black group, three models of Spearman's hypothesis were fitted. These were the same strong, contra, and weak models used above. The contra model can be thought of as a model nested within the weak model, which involved restricting the spatial factor for this sample. The contra model simply adds a restriction to g. Resultantly, the CFI for the contra model was 0.752, RMSEA was 0.121, and the Mc was 0.336. The weak model fared much better with a CFI of 0.991, RMSEA of 0.040, and Mc of 0.953. However, perhaps due to the small number of tests and their wide diversity, the strong model, in which all non-*g* factors were constrained to zero, fit best, and in fact, fit better than any other model, with a CFI of 0.999, RMSEA of 0.018, χ2 /df of 3.86, and Mc of 0.968. This makes the results of Kirkegaard et al. [18] very appropriate, since they used a *g* score, and the best-fitting model here only involved *g*.

The weak model for the Hispanic group was fitted first and it was fitted based on the verbal factor, which was constrained to zero. The resultant ΔCFI was -0.01, while ΔRMSEA was 0.09 and ΔMc was −0.078. The contra model, in which *g* was also constrained to zero, had a ΔCFI −0.013, a ΔMc of -0.037, and a ΔRMSEA of 0.013. The strong model, in which all group factors were constrained, had the best fit with a ΔCFI of −0.009, ΔMc of −0.027, and ΔRMSEA of 0.08. In this group, as in the last, the strong form of Spearman's hypothesis seemed to show the best fit, while the contra appeared to be the worst with every parameter non-invariant and the weak showed a borderline fit. It is quite possible that the young age of the sample, the small number of tests, or the diverse composition of the relatively small battery play a role in the acceptance of the strong form of SH in these groups.

To summarize the above, we found that the weak form of Spearman's hypothesis holds for the AA-EA comparison in the TCP dataset. What's more, we have found that the strong form of Spearman's hypothesis holds in the PING dataset when comparing HAs and AAs to EAs, though the reason for this is uncertain. This implies that Spearman's hypothesis has been confirmed with MGCFA three times here (in the TCP once and in the PING twice). In addition to these findings, we have found that all of these batteries display SFI and as such are unbiased cognitive assessments.

1.5. Assessing SES Effects with MGCFA in the PCNB

We tested how large of a biasing factor SES could be when modeled as a background variable in the PCNB homogeneous latent variances MGCFA model. When SES was modeled as affecting only the subtests, it reduced mean racial differences in *g* by 0.166 *d* and turned the differences in executive function and episodic memory insignificant. At the same time, this increased the difference in complex cognition by 0.548 *d*. SES had a significant effect on all subtests. Next, a model in which SES affected all subtests and latent factors was fitted and it was found that this model had insignificantly worse fit and latent differences compared to the other model, though SES no longer significantly affected the PFMT, LNB, PCET, VOLT, or PWMT subtests and it did not significantly affect the episodic memory factor directly. A model in which SES directly affected only latent abilities was fitted and had substantially worse fit (e.g., ΔCFI = 0.011 and ΔRMSEA = 0.012), though, to note, it only made the difference in episodic memory insignificant and only significantly changed the difference in complex cognition, again, increasing the white advantage. Fitting the model in which SES affected subtests and latent abilities onto the weak SH model in which episodic memory was restricted led to no significant decrement in fit and reduced the difference in *g* by 0.116 *d* whilst making the difference in executive function insignificant and again increasing the difference in complex cognition, this time by 0.513 *d*. Only the path from SES to the PCPT was insignificant in this model. Overall, SES modeled in MGCFA only reduced the latent racial differences by a small amount.

Next, we fitted a MGCFA model in which European admixture was a background variable affecting subtests and latent factors, like the best-fitting SES model. The inclusion of European admixture as a background variable increased the difference due to *g* by 0.24 *d* whilst decreasing the differences in complex cognition by 0.128 *d* and increasing the differences in episodic memory by 0.607 *d* and making the difference in executive functioning insignificant. When this was fitted to the weak Spearman's hypothesis model, all paths remained significant, but the difference in *g* was only slightly reduced by 0.062 *d*, whilst the difference in complex cognition increased by 0.349 *d* and the executive functioning gap was rendered insignificant. Next, modeling both variables simultaneously reduced the *g* gap by 0.188 *d* while rendering the gaps in episodic memory and executive functioning insignificant and increasing the complex cognition gap by 0.428 *d*. SES only significantly affected the WRAT, PVRT, PCPT, PMAT, PWMT, *g*, and complex cognition while the effects of European admixture were only significant for the PVRT, PLOT, *g*, complex cognition, and executive functioning. In the weak SH model, the difference in *g* was reduced by 0.259 *d*, whilst the difference in complex cognition increased by 0.599 *d* and the difference in executive functioning was rendered insignificant. In this final model, SES significantly affected the WRAT, PVRT, PCPT, LNB, g, complex cognition, and executive function whilst European admixture affected the PVRT, PCET, PLOT, g, complex cognition, and executive functioning.

Overall, SES and European admixture modeled as background variables moderately affected the mean differences in the latent variables in the PCNB. It is possible that additional measures of SES or, better yet, a latent SES factor, could have offered more substantial mediation, but we were unfortunately unable to assess this possibility.

Section 2: Color in Sibling Pairs

In order to assess whether the effect of phenotype-based discrimination can account for the observed differences in part or in whole, it's not enough to declare that the effect of skin color overall is minor or that its inclusion as an endogenous covariate in a regression leads to little change and no significant effect. In point of fact, in a regression of X on Y, the inclusion of correlated downstream variable Z as an endogenous covariate will still reduce the β for X despite being uninvolved in the relationship between X and Y. Similarly, a variable such as education can reduce a real main effect of, say, IQ on income if it subsumes the variance from IQ and additional income-relevant personality and opportunity variables, even if it has no actual impact on income.

It's inappropriate to conclude from our own data that skin color does not mediate the association between ancestry and intelligence, as colorist theorists would suggest, because in the population at large, ancestry and skin color are confounded; intelligence and skin color are thus also confounded. However, in full sibling pairs, color and ancestry are not related due to random segregation and it would be strange for the relationship between intelligence and skin color to be pleiotropic because of the differing complexities of the genetic architectures of the two traits (intelligence is extremely complex and skin color is not). A sibling control should, therefore, elucidate whether the relationship between skin color and intelligence is a causal one.

Since the TCP does not supply sibling information we had to determine sibships by identity-by-descent in PLINK. Full siblings and dizygotic twins were determined to be those between 38 and 62% IBD since this region of the IBD histogram was discontinuous with the lower (half-sibling and below) and higher (monozygotic twins) regions. Only black siblings were used since colorism does not generally predict an effect within white ethnicities. There were only 108 black full siblings to use after removing individuals with the correct amount of IBD but different family background variables. The

SD for color was 5.05, which was not much less than the full black sample. The relationship between color and intelligence within these pairs was an insignificant r = 0.01 (p = 0.92). This was itself significantly different from the relationship in the full black, mixed, and white sample and the relationship in the combined black and mixed samples, but not from the relationship in the only full black sample (p = 0.35) or from the relationship in the whole sample of black siblings (i.e., the sibling average; r = −0.065, itself not significant). On the other hand, the relationship between admixture and intelligence across sibling pairs was r = 0.071 and within them r = 0.059, neither of which is significantly different from zero or the relationship in the full black sample. These data should be incorporated into a future meta-analysis of the effect within sibling pairs. Incidentally, the effect of ancestry is significantly different from zero but not from the overall group figure, whereas the effect of color is just not significantly different from zero in a sample of all full siblings (black, mixed, and white) in the dataset.

This analysis is inherently limited due to the small sample size and the fact that our measure of skin color is not actually a measure of skin color, but a genetic proxy. In order to assess whether results reached using this proxy are valid, we would need to replicate this analysis in a dataset containing both genetic data and skin color information. Since this was not available, not only these, but all of the results involving color will have to be considered tentative. Nonetheless, if a future analysis supports the tenability of this measure, the result should be considered for use in a future meta-analysis. It may even be useful to pool case-level data from siblings using this common measure in the future.

Section 3: Additional Admixture Plots

Below, we provide additional admixture plots. These show the relation between cognitive ability and either European ancestry or skin color.

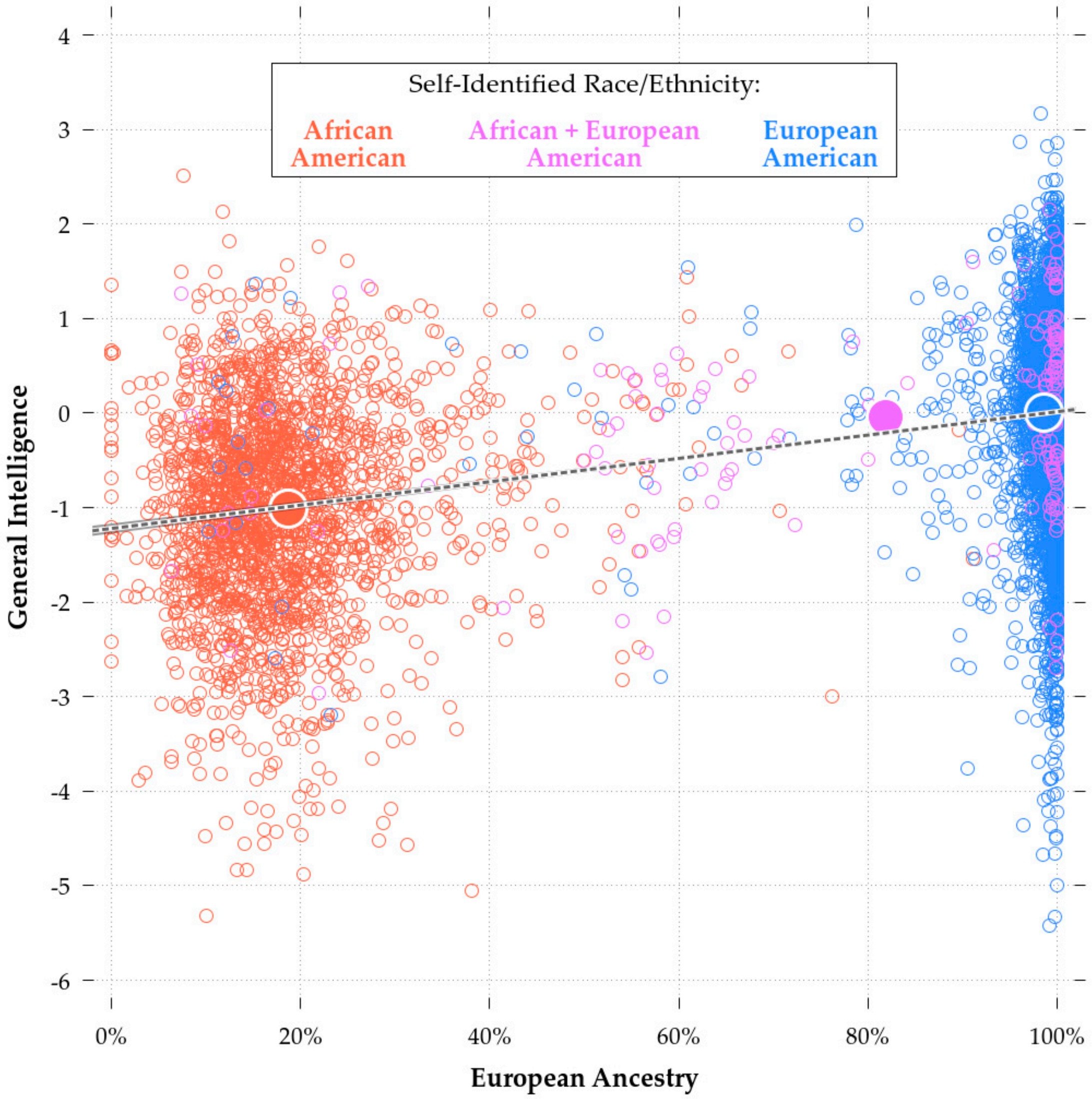

Figure S1. The Relationship between g and European Ancestry with Group Means.

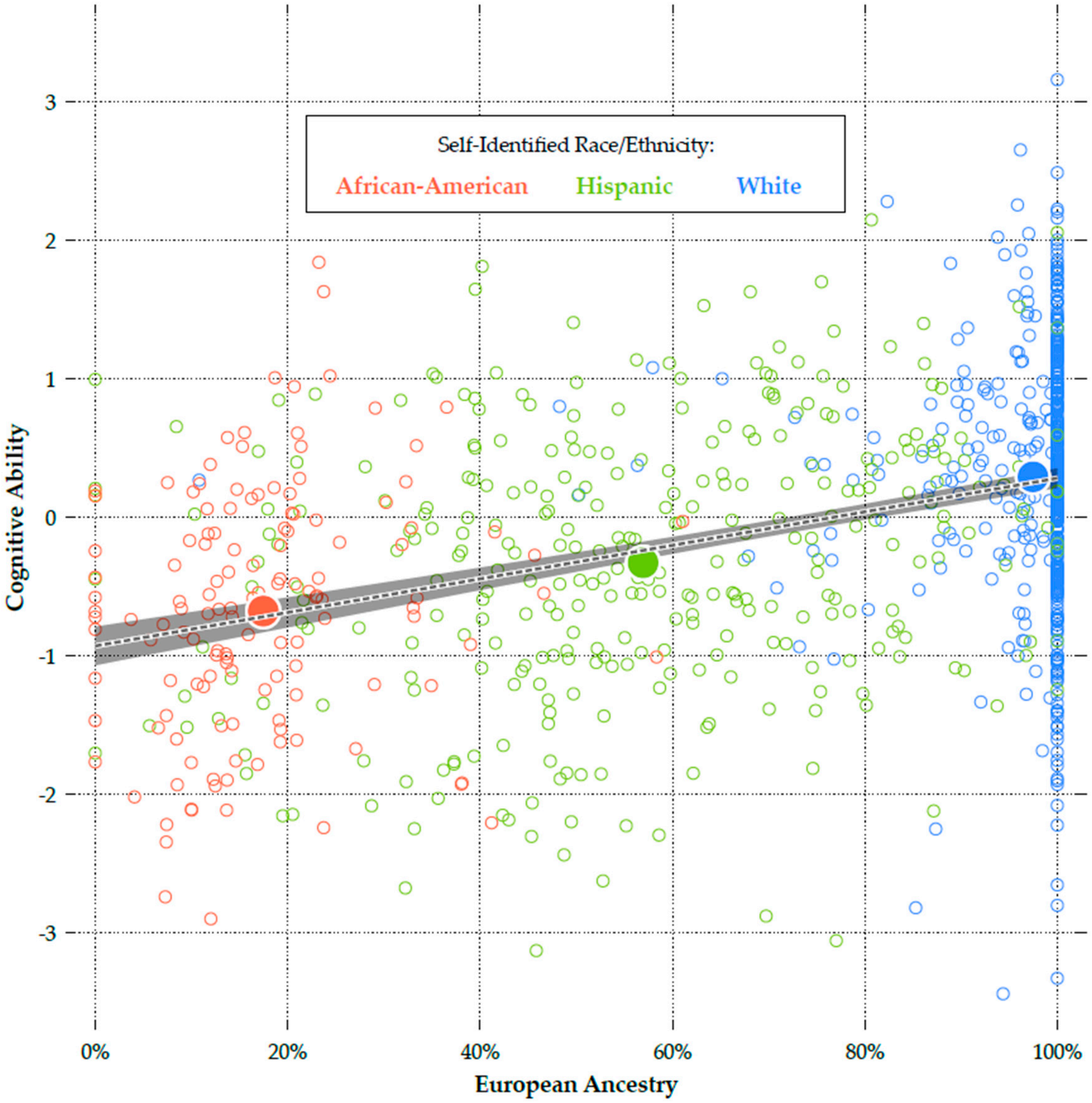

Figure S2. The Relationship between g and European Ancestry with Group Means in the PING Sample.

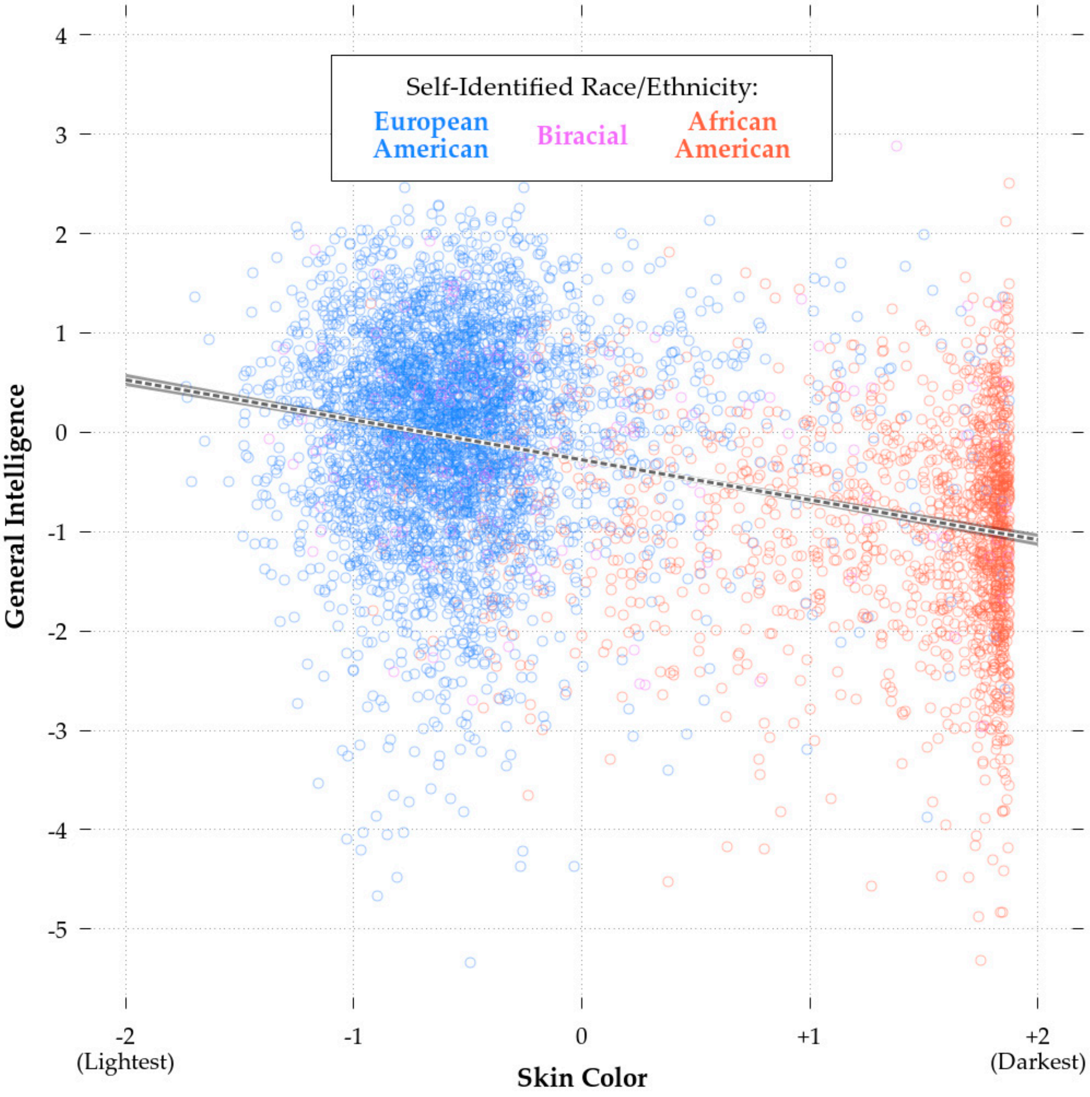

Figure S3. The Relationship between g and Color Scores.

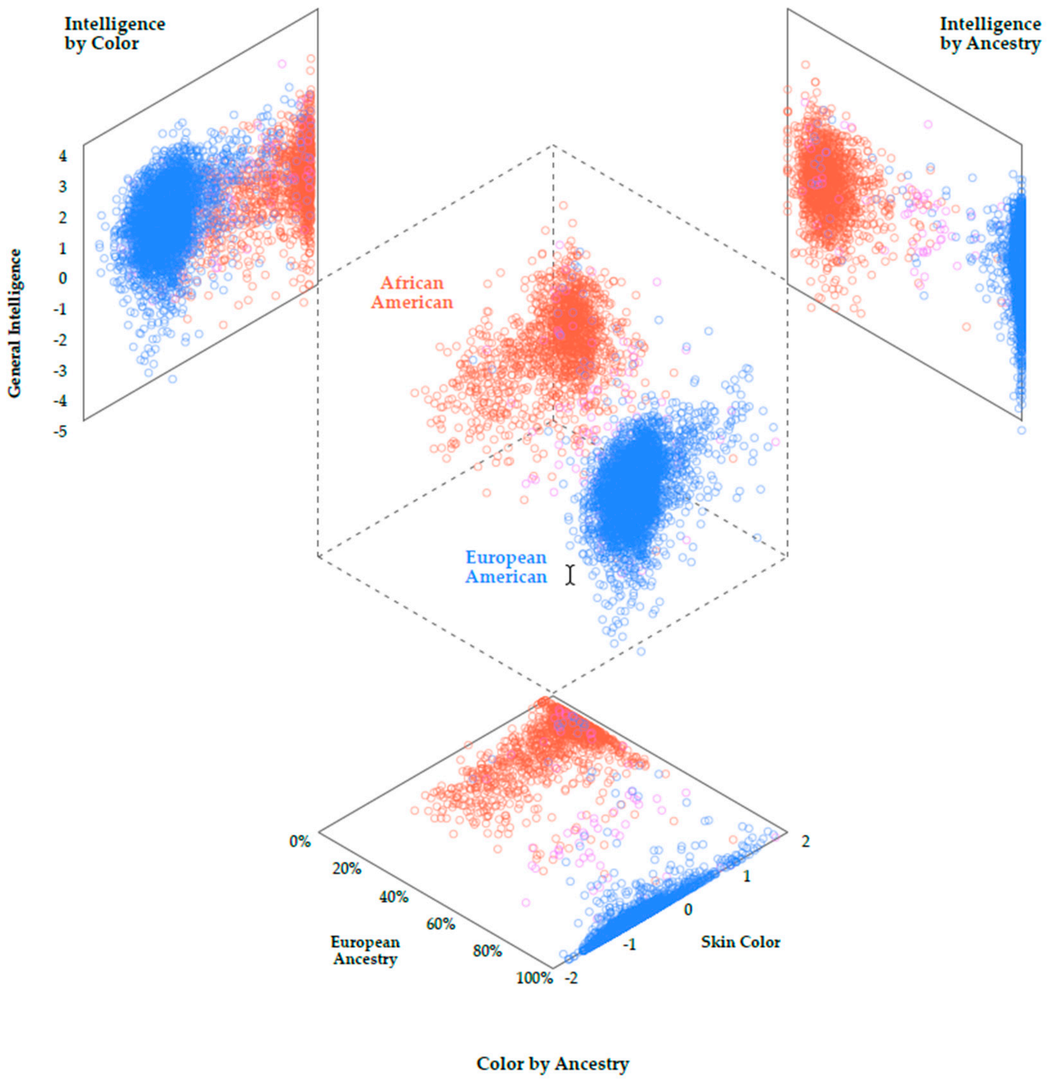

Figure S4. Three-Dimensional Plot of Color, Admixture, and General Intelligence.

References

1. Jensen, A.R. The nature of the black–white difference on various psychometric tests: Spearman's hypothesis. *Behav. Brain Sci.* **1985**, *8*, 193–219.
2. Jensen, A.R. The *g* Factor: The Science of Mental Ability. In *The g Factor: The Science of Mental Ability*. Praeger Publishers/Greenwood Publishing Group: Westport, CT, USA, 1998.
3. Jensen, A.R. "Spearman's hypothesis." In *Intelligence and Personality: Bridging the Gap in Theory and Measurement*, S. Messick, J.M., Collis Eds.; Lawrence Erlbaum Associates Publishers: Mahwah, NJ, USA, 2001.
4. te Nijenhuis, J.; Choi, Y.Y.; van den Hoek, M.; Valueva, E.; Lee, K.H. Spearman's hypothesis tested comparing Korean young adults with various other groups of young adults on the items of the Advanced Progressive Matrices. *J. Biosoc. Sci.* **2019**, 1–38, doi: 10.1017/S0021932019000026.
5. Rushton, J.P. The "Jensen effect" and the "Spearman-Jensen hypothesis" of Black-White IQ differences. *Intelligence* **1998**, *26*, 217–225.
6. Rushton, J.P.; Jensen, A.R. The rise and fall of the Flynn Effect as a reason to expect a narrowing of the Black–White IQ gap. *Intelligence* **2010**, *38*, 213–219.
7. Dolan, C.V.; Hamaker, E.L. (2001). *Investigating Black-White Differences in Psychometric IQ: Multi-Group Confirmatory Factor Analyses of The Wisc-R and Kabc and A Critique of The Method of Correlated Vectors*. Nova Science Publishers: Hauppauge, NY, USA, 2001.

8. Wicherts, J.M. Psychometric problems with the method of correlated vectors applied to item scores (including some nonsensical results). *Intelligence* **2017**, *60*, 26–38.
9. Dolan, C.V. Investigating Spearman's Hypothesis by Means of Multi-Group Confirmatory Factor Analysis. *Multivar. Behav. Res.* **2000**, *35*, 21–50.
10. Frisby, C.L.; Beaujean, A.A. Testing Spearman's hypotheses using a bi-factor model with WAIS-IV/WMS-IV standardization data. *Intelligence* **2015**, *51*, 79–97.
11. Lubke, G.H.; Dolan, C.V.; Kelderman, H. Investigating Group Differences on Cognitive Tests Using Spearman's Hypothesis: An Evaluation of Jensen's Method. *Multivar. Behav. Res.* **2001**, *36*, 299–324
12. Mellenbergh, G.J. Item bias and item response theory. *Int. J. Educ. Res.*, **1989**, *13*, 127–143.
13. Adolf, J.; Schuurman, N.K.; Borkenau, P.; Borsboom, D.; & Dolan, C.V. Measurement invariance within and between individuals: a distinct problem in testing the equivalence of intra- and inter-individual model structures. *Front. Psychol.* **2014**, *5*, doi: 10.3389/fpsyg.2014.00883.
14. Jak, S.; Jorgensen, T.D. Relating Measurement Invariance, Cross-Level Invariance, and Multilevel Reliability. *Front. Psychol.* **2017**, *8*, 1640.
15. Lubke, G.H.; Dolan, C.V.; Kelderman, H.; Mellenbergh, G.J. On the relationship between sources of within- and between-group differences and measurement invariance in the common factor model. *Intelligence*, **2003**, *31*, 543–566.
16. Beaujean, A.A. *Latent Variable Modeling Using R : A Step-by-Step Guide*. Routledge/Taylor & Francis Group: New York, NY, USA, 2014.
17. Barrett, P. Structural equation modelling: Adjudging model fit. *Personal. Individ. Differ.* **2014**, *42*, 815–824.
18. Schermelleh-Engel, K.; Moosbrugger, H.; Müller, H. Evaluating the Fit of Structural Equation Models: Tests of Significance and Descriptive Goodness-of-Fit Measures. *Methods Psycholo. Res.* **2003**, *8*, 23–74.
19. Moore, T.M.; Reise, S.P.; Gur, R.E.; Hakonarson, H.; Gur, R.C. Psychometric Properties of the Penn Computerized Neurocognitive Battery. *Neuropsychology* **2015**, *29*, 235–246.
20. Hu, L.; Bentler, P.M. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Struct. Equ. Model. Multidiscip. J.* **1999**, *6*, 1–55.
21. Cheung, G.W.; Rensvold, R.B. Evaluating Goodness-of-Fit Indexes for Testing Measurement Invariance. *Struct. Equ. Model. Multidiscip. J.* **2002**, *9*, 233–255.
22. Chen, F.F. Sensitivity of Goodness of Fit Indexes to Lack of Measurement Invariance. *Struct. Equ. Model. Multidiscip. J.* **2007**, *14*, 464–504
23. Khojasteh, J.; Lo, W.J. Investigating the Sensitivity of Goodness-of-Fit Indices to Detect Measurement Invariance in a Bifactor Model. *Struct. Equ. Model. Multidiscip. J.* **2015**, *22*, 531–541.
24. Putnick, D.L.; Bornstein, M.H. Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Dev. Rev.* **2016**, *41*, 71–90.
25. line, R.B. (2012). Assumptions of structural equation modeling. In Handbook of structural equation modeling, R. Hoyle Ed.; Guilford Press: New York City, NY, USA, 2012; pp. 111–125
26. West, S.G.; Finch, J.F.; Curran, P.J. (1995) Structural Equation Models with Non Normal Variables: Problems and remedies. In *Structural Equation Modeling: Concepts, Issues, and Applications*, R.H., Hoyle Ed.; SAGE Publications: Thousand Oaks, CA, USA, 1995, PP. 56–75.
27. Dixon, W.J. Efficient Analysis of Experimental Observations. *Annu. Rev. Pharmacol. Toxicol.* **1980**, *20*, 441–462.
28. Beaujean, A. John Carroll's views on intelligence: Bi-factor vs. higher-order models. *J. Intell.* **2015**, 3, 121–136.
29. Merkle, E.; Rosseel, Y.; Garnier-Villarreal, M.; Jorgensen, T.D.; Hoofs, H.; Schoot, R. van de. blavaan: Bayesian Latent Variable Analysis (Version 0.3-4). Available online: https://rdrr.io/cran/blavaan/ (accessed on 29 August 2019)
30. van de Schoot, R.; Kluytmans, A.; Tummers, L.; Lugtig, P.; Hox, J.; Muthén, B. Facing off with Scylla and Charybdis: a comparison of scalar, partial, and the novel possibility of approximate measurement invariance. *Front. Psychol.* **2013**, *4*, doi: 10.3389/fpsyg.2013.00770.
31. Bartholomew, D.J. Measuring intelligence: Facts and fallacies. In *Measuring Intelligence: Facts and Fallacies*. Cambridge University Press: New York, NY, USA, 2004.
32. Mollon, J.; Knowles, E.E.M.; Mathias, S.R.; Gur, R.; Peralta, J.M.; Weiner, D.J.; Robinson, E.B.; Gur, R.E.; Blangero, J.; Almasy, L.; et al. Genetic influence on cognitive development between childhood and adulthood. *Mol. Psychiatry* 2018, 1, 1.

33. Kirkegaard, E.O.W.; Woodley of Menie, M.A.; Williams, R.L.; Fuerst, J.; Meisenberg, G. Biogeographic Ancestry, Cognitive Ability and Socioeconomic Outcomes. *Psych*, **2019**, *1*, 1–25.
34. Wicherts, J.M.; Dolan, C.V. Measurement Invariance in Confirmatory Factor Analysis: An Illustration Using IQ Test Performance of Minorities. *Educ. Meas. Issues Pr.* **2010**, *29*, 39–47.