Supplementary File 1 for "More Research Needed: There is a Robust Causal vs. Confounding Problem for Intelligence-associated Polygenic Scores in Context to Admixed American Populations":

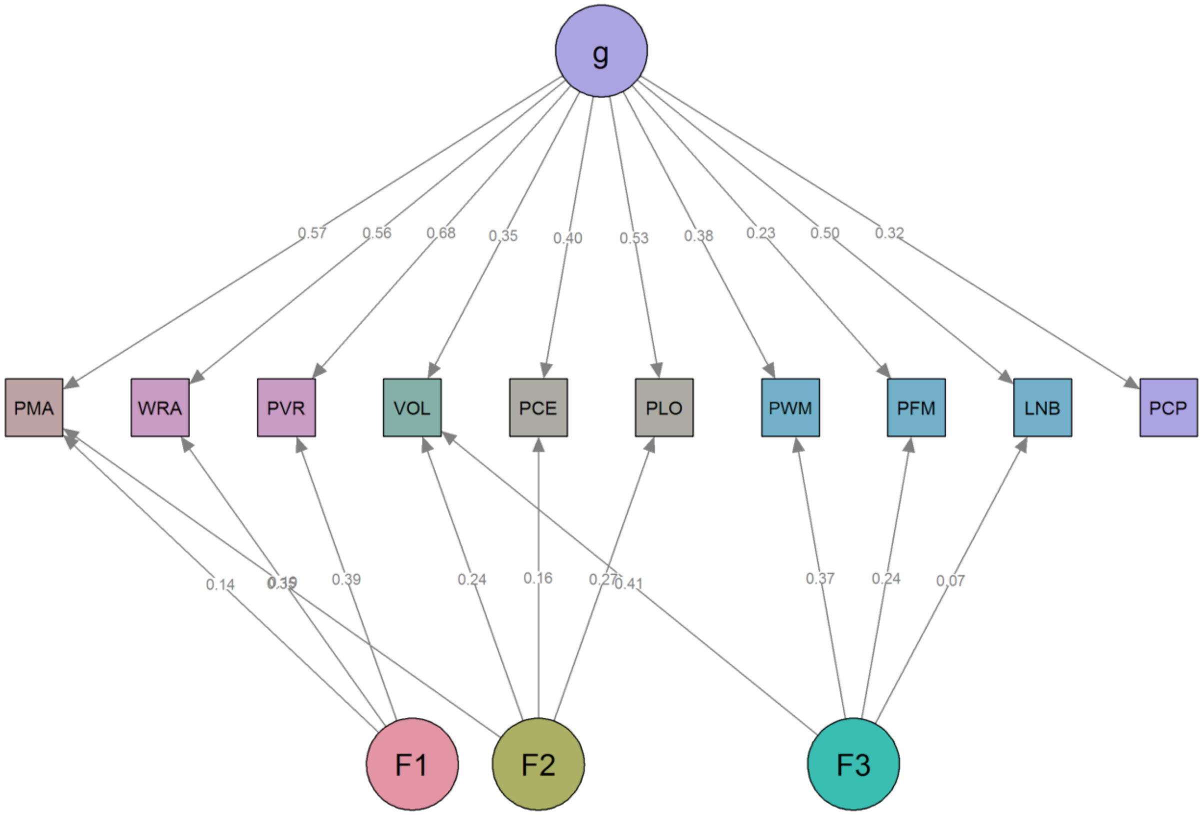
1. Psychometric Assessment of the Penn Computerized Neurocognitive Battery

The 14 PCNB subtests were designed to measure five broad behavioral domains: Executive Control, Episodic Memory, Complex Cognition, Social Cognition, and Sensorimotor Speed. The subtests are as follows: 1. Executive Control: Penn Conditional Exclusion Test (PCET; meant to assess Mental Flexibility), Penn Continuous Performance Test (PCPT; Attention), and Letter N-Back Task (LNB; Working Memory); 2. Episodic Memory: Penn Word Memory Task (PWMT; Verbal Memory), Penn Face Memory Task (PFMT; Face Memory), and Visual Object Learning Test (VOLT; Spatial Memory); 3. Complex Cognition: Penn Verbal Reasoning Test (PVRT; Language Reasoning), Penn Matrix Reasoning Test (PMRT; Nonverbal Reasoning), and Penn Line Orientation Test (PLOT; Spatial Ability); 4. Social Cognition: Penn Emotion Identification Test (PEIT; Emotion Identification), Penn Emotion Differentiation Test (PEDT; Emotion Differentiation), and Penn Age Differentiation Test (PADT; Age Differentiation). 5. Sensorimotor Speed: Motor Praxis Test (MP; Sensorimotor Speed), and Finger Tapping (Tap; Sensorimotor Speed) (Lasker et al., 2019). Participants additionally completed the Wide Range Achievement Test (WRAT), a highly-reliable broad ability measure (Moore et al., 2015).

We excluded the approximately 1.5% of individuals for whom data were missing for at least half the tests. We then imputed values for the remaining cases using IRMI (iterative robust model-based imputation; Templ, Kowarik, & Filzmoser. 2011; Templ, Kowarik, Alfons, & Prantner, 2019). The effects of age and sex on subtest scores have previously been detailed (Gur et al., 2012; Roalf et al., 2014). To handle non-linear effects, we residualized the subtest variables for age and sex using a natural (i.e., restricted cubic) spline model before performing factor analysis.

We ran both exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). The model with Social Cognition and the Sensorimotor Speed factors exhibited a lack of indicator coherence and did not converge. As such, we removed the five subtests related to these two factors and ran EFA/CFA for the 10 remaining ones. A model, based on the remaining tests, with Complex Cognition, Executive Control, and Episodic Memory factors as specified by Moore et al.'s (2015) five factor model also did not converge. However, we identified a similar three factor model (with analogous Executive Control, Episodic Memory, Complex Cognition broad factors), which had good fit among European and Hispanic Americans. This is shown in Figure S1.

Figure S1. Three Factor Model for the Penn Cognitive Battery.



This same bifactor model also had a good fit among European and African Americans. To ensure that the measure was unbiased between our major ethnic groups, we performed Multi-group confirmatory factor analyses (MGCFA) to determine if Measurement Invariance (MI) held for European and African Americans and for the European and Hispanic Americans (Dolan & Hamaker, 2001; Lubke, Dolan, Kelderman, & Mellenbergh, 2003). Results for European and African Americans were not interpretively different from those reported by Lasker et al. (2019), despite, in this case, using Full Information Maximum Likelihood (FIML) instead of listwise deletion to handle missing subtest scores. Results for European and Hispanic Americans are discussed below.

2. Assessment of Measurement Invariance

Multi-group confirmatory factor analysis (MGCFA) is a technique often used to assess the measurement invariance (MI) of different psychological assessments including intelligence tests (Dolan, 2000; Dolan & Hamaker, 2001; Lubke, Dolan & Kelderman, 2001; Frisby & Beaujean, 2015). When MI holds, the assessment being examined is generally considered unbiased in the groups being compared. Resultantly, MI is taken to imply that the constructs measured in both groups are alike. MI is assessed by fitting a series of increasingly restrictive models and analyzing model fit at each step (van de Schoot, Lugtig & Hox, 2012). We assessed MI for Hispanic and European American comparisons with the Penn Computerized Neurocognitive Battery (PCNB).

Model fit was assessed with multiple indices. We adopted the same procedure, model, and criteria as in Lasker, Pesta, Fuerst, & Kirkegaard (2019), who examined MI for non-Hispanic European Americans and non-Hispanic African Americans. Specifically, for assessing measurement invariance, we adopted Cheung and Rensvold's (2002) and Chen's (2007) frequently accepted criteria. Specifically, we regarded 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 was untenable. See Putnick & Bornstein (2016) for a review of common conventional criteria.

For this analysis, we ran MGCFA on the set of individuals with sufficient cognitive data for imputation (506 Hispanic and 4,914 European Americans) without limiting the analysis to only those who passed quality controls for computing genetic ancestry. This set includes individuals with imputed subtest scores (when fewer than half of the subtests were missing). Our results are presented in tables S1-S3 and the lavaan model syntax is provided at the end of this supplement.

Table S1. Bifactor Solution for Hispanic and European Americans on the Philadelphia Computerized Neurocognitive Battery.

Model	MI Step	χ2	Df	CFI	ΔCFI	RMSEA	ΔRMSEA	Mc	ΔMc	SRMR
1	Configural	115.45	48	0.992	-	0.023	-	0.994	-	0.013
2	Metric	135.60	65	0.991	-0.001	0.020	-0.003	0.994	0	0.015
3	Scalar	171.63	71	0.988	-0.003	0.023	0.003	0.991	-0.003	0.017
4	Strict	252.43	81	0.979	-0.009	0.028	0.005	0.984	-0.007	0.020
5	Latent Variances	252.43	81	0.979	0	0.028	0	0.984	0	0.020
6	Means	376.71	85	0.964	-0.015	0.036	0.008	0.973	0.011	0.029
5a	Strong	268.94	84	0.977	-0.002	0.029	0.001	0.983	-0.001	0.021
5b	Weak	254.38	82	0.979	0	0.028	0	0.984	0	0.021
5c	Contra	374.09	83	0.964	-0.015	0.036	0.008	0.974	-0.010	0.029

Note: Combined *N* = 5,420, with 506 Hispanic Americans and 4,914 European Americans. The latent variance model merely changes the identification constraint to the variances from a single loading for each modeled factor.

In Table S1, Models 5a to 5c further assess Spearman's hypothesis (Jensen, 1998; Frisby & Beaujean, 2015), which is more fully discussed by Lasker et al. (2019). The strong model (5a) leaves only *g* to vary between groups. The weak model (5b) leaves *g* to vary and constrains complex cognition; it should be noted that it was possible to constrain any set of the broad factors without a meaningful decrease in model fit (ΔCFI for weak models ranged from 0 to -0.001 out of all six possible models). The contra model (5c) constrains *g* and carries the broad factor constraints from the weak model. Contra model fits were always worse (approximately and absolutely) than the fits of comparable weak models (ΔCFI ranged from -0.005 to -0.015). 5a-c are each compared to model 5. Neither the strong nor weak model fits worse than the model with latent variances constrained in terms of approximate fit; however, using a χ2 test, the weak model does not fit worse while the strong model does (like the contra model). The fit for the chosen contra model could be rejected with a ΔCFI of <-0.01 accompanied by a notably elevated χ2. These results tentatively support either the weak or strong model over the contra model with approximate fits and absolutely support the weak model with a χ2 test.

Tables S2 and S3 show, respectively, the standardized mean differences based on the model with constrained latent variances and the weak Spearman’s hypothesis model. In the weak

Spearman's hypothesis model, the European-Hispanic American difference in *g* is 0.668 Hedge's *g* (positive values favor European Americans and vice-versa). There are also small to moderate differences in executive functioning (Hedge's *g* = -0.342) and episodic memory (Hedge's *g* = -0.234) net of *g* which favor Hispanics. In the contra or baseline models, broad factors more strongly favor European Americans, as the differences associated with *g* in the latent variances or weak models are distributed among the other factors in the absence of *g*.

The values of ωh and ωt for this battery were 0.69 and 0.77 respectively; 90% of the reliable variance was thus attributable to *g*. The ECV for *g* was 70%, PUC was 0.78, and H was 0.76, with these values being uniformly too low for complex cognition (ECV = 9%, H = 0.27), executive functioning (9%, 0.24%), and episodic memory (12%, 0.31). Using the method from Dolan (2001), in the latent variances model, an average of 67% of the between-group differences were accounted for by *g*; 65% of the differences in the indicators for complex cognition, 62% for executive functioning, and 58% for episodic memory.

In the selected weak Spearman's hypothesis model, 73% of the group differences are accounted for by *g* and the proportion of the differences in the indicators for executive functioning were unchanged. The correlation between the vector of group differences and the vector of *g* loadings is *r* = 0.524. This same correlation for the complex cognition, executive functioning, and episodic memory loadings are, respectively, *r* = -0.420, *r* = 0.113, *r* = -0.198, and overall, *r* = -0.159. Values for Mardia's b1p and b2p were 16.403 and 132.853 (Mardia, 1980).

Table S2. Factor Score Differences between Hispanic and European Americans based on the Model with Constrained Latent Variances.

Factor	Estimate	SE	Lower 95% CI	Upper 95% CI
G	0.614	0.071	0.474	0.754

Complex Cognition	0.123	0.088	-0.049	0.296
Executive Functioning	-0.259	0.126	-0.506	-0.012
Episodic Memory	-0.192	0.078	-0.344	-0.040

Note: Positive values indicate higher European American scores and vice-versa. Estimates are in terms of Hedge’s *g*. Combined *N* = 5,454 with 515 Hispanic Americans and 4,939 European Americans.

Table S3. Factor Score Differences between Hispanic and European Americans based on the Weak Spearman’s Hypothesis Model.

Factor	Estimate	SE	Lower 95% CI	Upper 95% CI
G	0.668	0.064	0.543	0.793
Complex Cognition	0	-	-	-
Executive Functioning	-0.342	0.124	-0.585	-0.099
Episodic Memory	-0.234	0.075	-0.382	-0.087

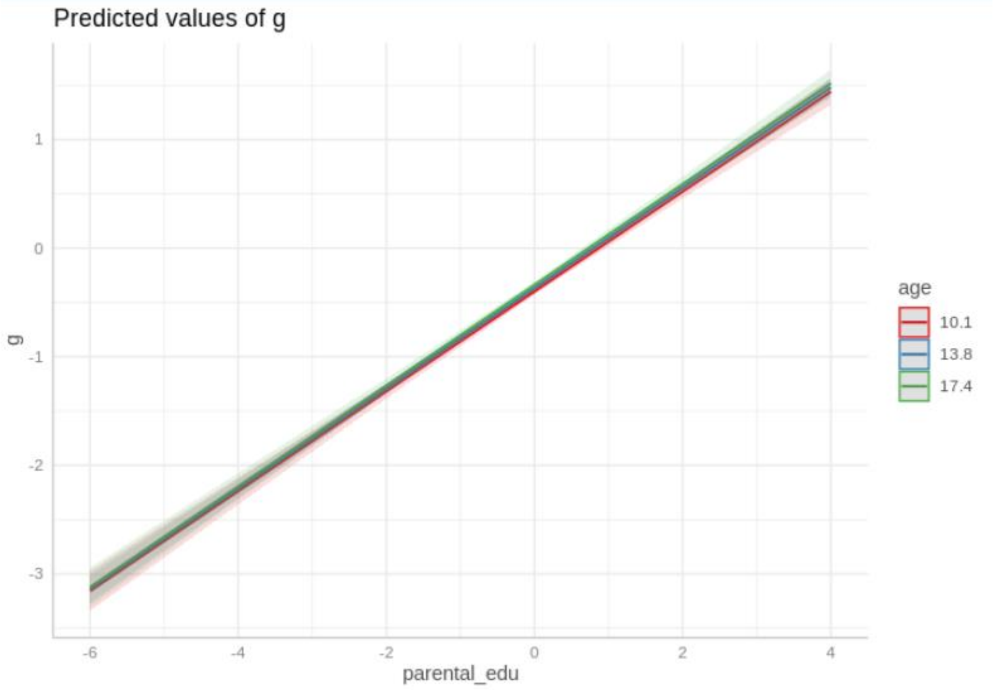
Note: Positive values indicate higher European American scores and vice-versa. Estimates are in terms of Hedge’s *g*. Combined *N* = 5,454 with 515 Hispanic Americans and 4,939 European Americans.

Generally, full factorial invariance held (by the conventional metrics cited above). Additionally, the weak form of Spearman’s hypothesis (Jensen, 1998), in which *g* accounts for the majority of variance in subtest scores, did not fit worse than baseline. In this model, 67-73% of the between-group differences are accounted for by *g* (model depending). In contrast, the contra models, in which the majority of subtest score differences were not attributable to *g*, and, arguably, the strong model, in which *g* accounts for all of variance in subtest scores, fit worse. Since it is often difficult to distinguish between Spearman Hypothesis models (e.g., strong vs. various forms of weak SH), and since the magnitude of *g* differences depends on the specification, for replicability and less dependence on researcher choice, we used *g*-scores from the exploratory factor analysis. These correlated at *r* = .97 with scores from the MGCFA model above.

3. Assessment of Age Effects

As there was a relatively wide age range (8 to 22 years), on a reader's request, we additionally investigated the validity of these scores across age groups. To do this we used parental education as a predictor and *g*-scores as the criterion. We used education as a predictor, since this variable is a known correlate of *g*, eduPGS, and genetic ancestry, and also since most cases had this variable. For this analysis, we limited the analytic sample to those with ancestry, *g*, and education. In Figure S2, we show the regression plot for parental education and *g* by age group, with participants grouped into three age distributions (for purposes of illustration). As is evident, age grouping has little effect. Moreover, in a regression model with parental education predicting *g*, the interaction term for age had a trivial, albeit statistically significant, effect (ß = =.008; *p* = .006; *N* = 7,846), likely due to the large sample size. Generally, parental education predicts *g*-scores more or less equally well across the age distribution, suggesting that our natural spline model effectively captured the age-related effects on subtests.

Figure S2. Regression Plot for Parental Education and g by Three Age Groups.



Means and standard deviations for *g* and the subtests are presented in Table S4. These (also including results for the five psychometrically biased subtests) are provided for the four main groups.

Table S4. General Intelligence (g) and Subtest Means and Standard Deviations for European (EA), European-African (EA-AA), African (AA), and Hispanic (HI) Americans.

	EA		EA-AA		AA		HI	
g	0.00	1.01	-0.14	1.05	-1.01	1.07	-0.57	1.13
PLOT*	-0.01	1.00	-0.22	0.96	-0.71	0.96	-0.36	1.04
PCPT*	0.00	1.00	-0.04	1.03	-0.33	1.16	-0.26	1.26
PCET*	0.00	1.00	0.03	0.97	-0.45	1.08	-0.26	1.08
LNB*	0.00	1.00	-0.02	0.98	-0.47	1.18	-0.29	1.12
VOLT*	0.00	1.00	-0.09	0.96	-0.35	1.11	-0.27	1.06
TAP	0.00	1.00	0.04	1.07	-0.07	1.06	0.02	0.98
PMRT*	0.01	1.00	-0.06	1.05	-0.56	0.94	-0.25	0.98
MP	0.00	1.00	-0.16	1.10	-0.11	1.04	-0.07	1.18
PEDT	0.00	1.01	-0.06	0.95	-0.15	1.16	-0.15	1.09
PVRT*	0.00	1.00	-0.15	1.07	-0.98	1.13	-0.58	1.18
PEIT	-0.01	1.01	-0.02	1.07	-0.05	1.10	0.02	1.02
PFMT*	-0.01	1.00	0.10	1.03	-0.04	1.08	0.06	1.04
PADT	0.00	1.01	-0.06	0.93	0.00	1.12	-0.06	1.02
PWMT*	0.00	1.01	0.07	1.03	-0.17	1.23	-0.11	1.16
WRAT*	0.00	1.00	-0.13	1.08	-0.85	0.95	-0.48	1.05

Note: *Denotes that the subtests were used in computing *g* scores.

4. Subtest Means and Standard Deviations by Hispanic Subgroup

We additionally provide the means and standard deviations for the subtests by Hispanic subgroup, though these are not used in any analyses. Scores for all 15 subtests are provided in Table S5, with an asterisk placed next to the 10 for which measurement invariance was found to hold.

Table S5. Subtest Means and Standard Deviations by Hispanic Subgroup.

	HI		**HI_EA**		**HI_AA**		**HI_OT**		**Other**	
g	-0.57	1.13	-0.33	1.17	-0.84	0.98	-0.65	1.17	-0.39	1.13
PLOT*	-0.36	1.04	-0.17	1.05	-0.62	1.03	-0.40	1.03	-0.21	0.98
PCPT*	-0.26	1.26	-0.19	1.14	-0.29	1.32	-0.46	1.38	-0.46	1.17
PCET*	-0.26	1.08	-0.25	1.12	-0.43	1.04	-0.25	1.06	-0.09	1.08
LNB*	-0.29	1.12	-0.23	1.12	-0.42	1.19	-0.33	0.99	-0.15	1.13
VOLT*	-0.27	1.06	-0.08	1.00	-0.51	1.09	-0.33	1.10	-0.12	0.99
TAP	0.02	0.98	0.14	0.96	0.07	0.91	-0.07	1.09	-0.05	0.97
PMRT*	-0.25	0.98	-0.13	1.02	-0.39	0.89	-0.25	1.00	-0.22	1.02
MP	-0.07	1.18	-0.02	1.17	0.05	0.83	-0.31	1.65	-0.02	0.99
PEDT	-0.15	1.09	0.02	1.15	-0.21	1.17	-0.25	1.06	-0.13	0.95
PVRT*	-0.58	1.18	-0.35	1.23	-0.74	1.14	-0.67	1.28	-0.52	1.06
PEIT	0.02	1.02	-0.10	1.13	-0.01	0.91	0.12	1.13	0.08	0.91
PFMT*	0.06	1.04	-0.03	1.02	-0.02	1.03	0.32	0.98	-0.01	1.09
PADT	-0.06	1.02	0.03	1.04	-0.13	1.11	-0.09	1.00	-0.02	0.92
PWMT*	-0.11	1.16	0.06	0.99	-0.09	1.00	-0.37	1.47	-0.03	1.12
WRAT*	-0.48	1.05	-0.34	1.01	-0.75	1.04	-0.52	1.01	-0.25	1.07

*Note: *Denotes the subtests, from the 10-subtest measurement invariant model, used to compute g scores. HI_EA = Hispanic European, HI_AA = Hispanic African, HI_EA = Hispanic Other, and Other = any other also with Hispanic ethnicity marked.*

References.

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.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating Goodness-of-Fit Indexes for Testing Measurement Invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, 9(2), 233–255.

Dolan, C. V. (2000). Investigating Spearman’s Hypothesis by Means of Multi-Group Confirmatory Factor Analysis. *Multivariate Behavioral Research*, 35(1), 21–50.

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

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

Gur, R. C., Richard, J., Calkins, M. E., Chiavacci, R., Hansen, J. A., Bilker, W. B., ... & Abou-Sleiman, P. M. (2012). Age group and sex differences in performance on a computerized neurocognitive battery in children age 8− 21. *Neuropsychology*, 26(2), 251.

Jensen, A. R. (1998). *The g Factor: The Science of Mental Ability*. Westport, CT, US: Praeger Publishers/Greenwood Publishing Group.

Lasker, J., Pesta, B. J., Fuerst, J. G. R., & Kirkegaard, E. O. W. (2019). Global Ancestry and Cognitive Ability. *Psych*, 1(1), 431–459.

Lubke, G. H., Dolan, C. V., & Kelderman, H. (2001). Investigating Group Differences on Cognitive Tests Using Spearman's Hypothesis: An Evaluation of Jensen's Method. *Multivariate Behavioral Research*, 36(3), 299–324.

Mardia, K. V. (1980). Tests of unvariate and multivariate normality. In Analysis of Variance: Vol. 1. *Handbook of Statistics* (pp. 279–320).

Moore, T. M., Reise, S. P., Gur, R. E., Hakonarson, H., & Gur, R. C. (2015). Psychometric properties of the Penn Computerized Neurocognitive Battery. *Neuropsychology*, 29(2), 235.

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review*, 41, 71-90.

Roalf, D. R., Gur, R. E., Ruparel, K., Calkins, M. E., Satterthwaite, T. D., Bilker, W. B., ... & Gur, R. C. (2014). Within-individual variability in neurocognitive performance: Age-and sex-related differences in children and youths from ages 8 to 21. *Neuropsychology*, 28(4), 506.

van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology*, 9(4), 486–492.

Model Syntax :

```
F1 =~ WRAT + PVRT + PMAT

F2 =~ PMAT + PCET + VOLT + PLOT

F3 =~ VOLT + PWMT + PFMT + LNB
```

```
g =~ WRAT + PVRT + PMAT + PCET + VOLT + PLOT + LNB + PWMT + PFMT + PCPT
```

Supplementary File 2 for "More Research Needed: There is a Robust Causal vs. Confounding Problem for Intelligence-associated Polygenic Scores in Context to Admixed American Populations":

1. Detailed Discussion of the Color Variable.

We calculated phenotypic scores from genotypic data. We imputed phenotype based on genotype using the HIrisPlex-S web application (https://hirisplex.erasmusmc.nl/). HIrisPlex-S gives probabilities of The Fitzpatrick Scale skin type, which range from Type I "palest; freckles" to Type VI "deeply pigmented dark brown to darkest brown". The Fitzpatrick Scale skin types correspond with scores on Von Luschan's chromatic scale. For example, a Fitzpatrick Scale Type I classification corresponds with a Von Luschan's chromatic scale score of 0–6. This correspondence allowed us to transform the HIrisPlex-S skin type probabilities into a single, color measure. This was done by weighting the median score of each color type by the HIrisPlex-S predicted probability of each type. Owing to poor tagging of SNPs in the arrays, it was possible to compute color scores for only 3,862 European, 166 European-African, 1,557 African, and 398 Hispanic Americans.

The correlation between skin color and European ancestry for the combined sample was $r = -.87$ ($N = 6{,}050$), as shown in Figure S1. For Hispanics alone, the correlation was $r = -.67$ ($N = 398$), as shown in Figure S2. Note, the expected correlations are population specific, owing to differences in admixture range, admixture components, assortative mating, etc. (Kim, Edge, Goldberg, & Rosenberg, 2019). In this case, the estimate found for Philadelphian Hispanics is similar to those reported by others for similar populations (e.g., Puerto Rican: rho = .63; Parra, Kittles, & Shriver, 2004; R-square = .417 / $r = .65$; Bonilla, Shriver, Parra, Jones, & Fernández, 2004).

For the African American only (monoracial) sample, we are not aware of a directly comparable Philadelphia sample. However, previously, Scarr, Pakstis, Katz, & Barker (1977) reported a correlation of *r* = 0.21 and 0.27, respectively, between skin color and their ancestral and sample odds indexes of African ancestry, based on blood groups, in an African American sample. However, their ancestry-index likely had a reduced validity of around 50% (Jensen, 1982; Lasker et al. 2019). The corrected correlation would be around *r* = .4 to .5. This is similar to the relation found in a somewhat comparable sample (i.e., African Americans from Washington, D.C.: r_s = .44; Parra, Kittles, & Shriver, 2004). And both of these are comparable to the *r* = .39 we found for Philadelphian monoracial African Americans, so it is likely that our color estimates are reasonably precise.

Consistent with observations previously reported (e.g., Bonilla et al., 2004), the mean score for Hispanics was 22.19 which is intermediate to Type III (sometimes mild burn, tans uniformly) and Type IV (burns minimally, always tans well, moderate brown). Moreover, Hispanics who identified as European had a color score of 19.45, which was significantly darker than that for non-Hispanic European Americans at 14.70 (*t* (3,940) = 10.645). And, Hispanics who identified as African had a color score of 28.45, which was significantly lighter than that for non-Hispanic African Americans at 30.96 (*t* (1,671) = -4.393). Generally, these color values are consistent with known population values.

Figure S1. Regression Plot of the Relation Between Color (with Higher Values indicating Darker Color) and European Genetic Ancestry in the Combined Sample.

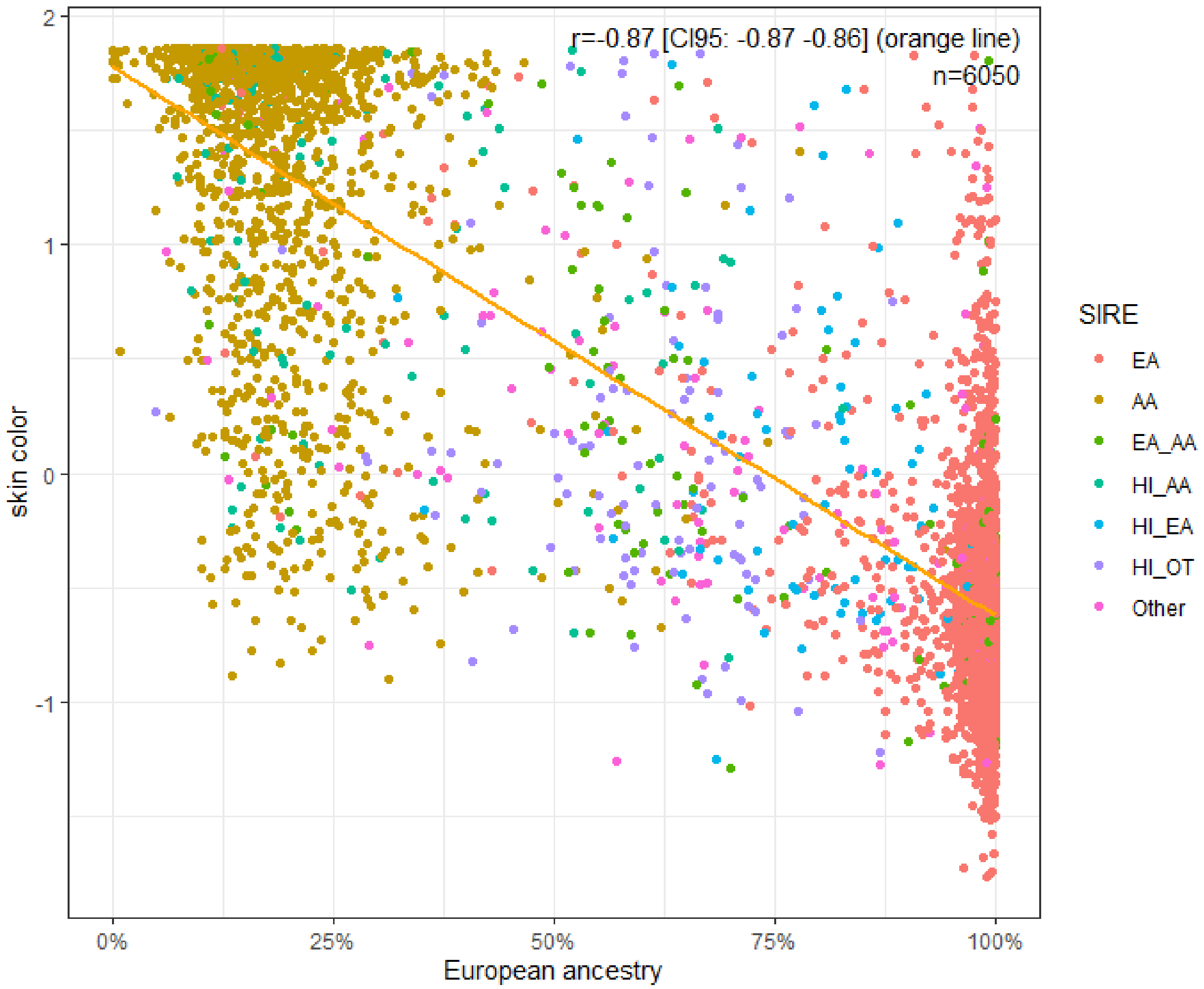
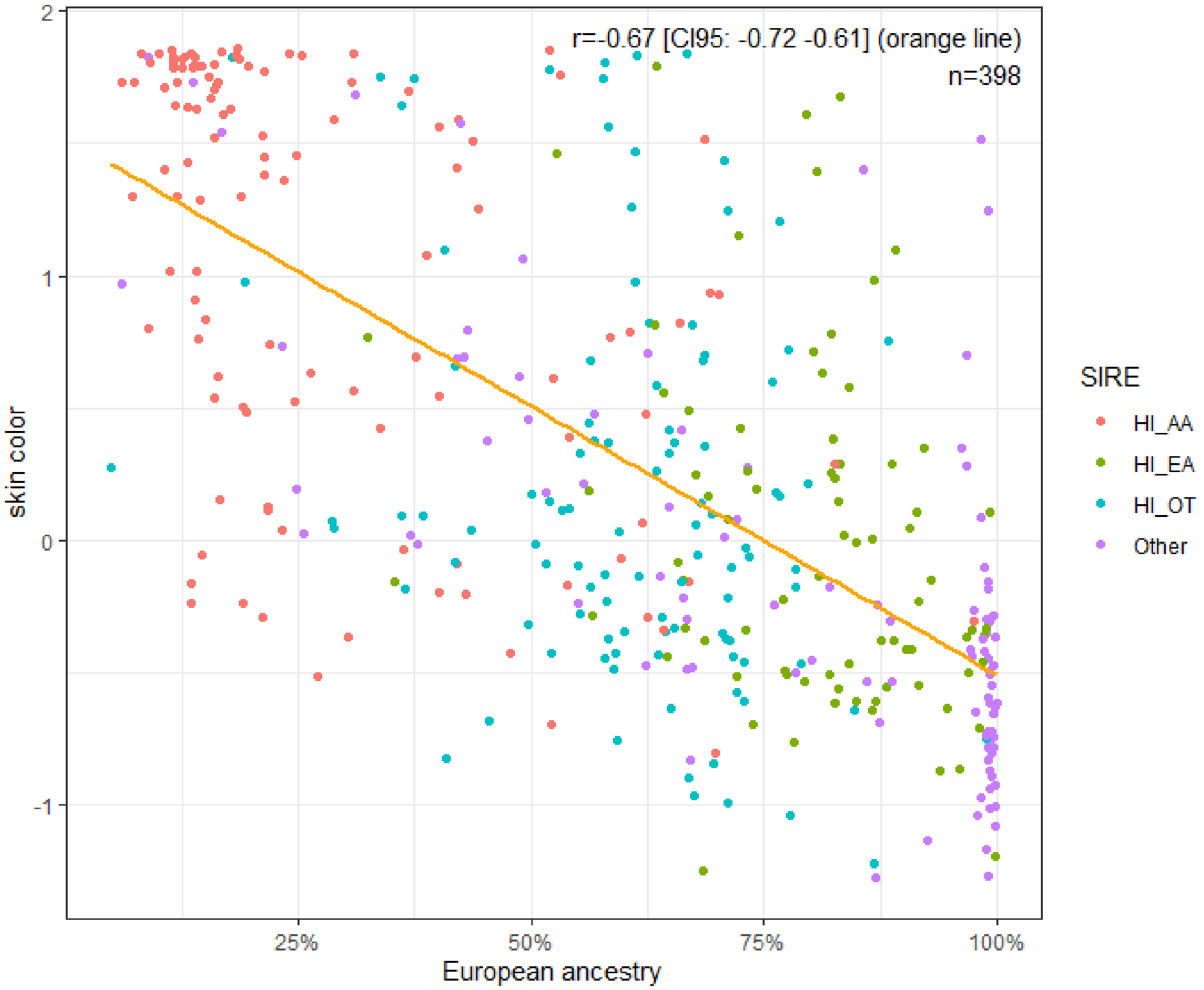


Figure S2. Regression Plot of the Relation Between Color (with Higher Values indicating Darker Color) and European Genetic Ancestry Among Hispanics.



References.

Bonilla, C., Shriver, M. D., Parra, E. J., Jones, A., & Fernández, J. R. (2004). Ancestral proportions and their association with skin pigmentation and bone mineral density in Puerto Rican women from New York city. *Human Genetics*, 115(1), 57-68.

Jensen A. R. (1981). Obstacles, Problems, and Pitfalls in Differential Psychology. In: Scarr, S. (1981). *Race, social class, and individual differences in IQ*. Hillsdale, NJ: Erlbaum.

Kim, J., Edge, M. D., Goldberg, A., & Rosenberg, N. A. (2019). Assortative mating and the dynamical decoupling of genetic admixture levels from phenotypes that differ between source populations. bioRxiv, 773663.

Lasker, J., Pesta, B. J., Fuerst, J. G., & Kirkegaard, E. O. (2019). Global Ancestry and Cognitive Ability. *Psych*, 1(1), 431-459.

Parra, E. J., Kittles, R. A., & Shriver, M. D. (2004). Implications of correlations between skin color and genetic ancestry for biomedical research. *Nature Genetics*, 36(11s), S54.

Scarr, S., Pakstis, A. J., Katz, S. H., & Barker, W. B. (1977). Absence of a relationship between degree of White ancestry and intellectual skills within a Black population. *Human Genetics*, 39(1), 69-86.

Supplementary File 3 for "More Research Needed: There is a Robust Causal vs. Confounding Problem for Intelligence-associated Polygenic Scores in Context to Admixed American Populations":

1. Bivariate Relationships Among the Variables for the Hispanic and for the Combined Group

The bivariate correlations allow comparison with effects sizes for the association between ancestry and SES reported previously (e.g., Kirkegaard, Wang, & Fuerst, 2017). Table S1 shows the correlations for Hispanics. Consistent with having a predominantly Puerto Rican sample (e.g., Via et al., 2011), Amerindian ancestry was only weakly (negatively) correlated with European ancestry. Cognitive ability and parental education were positively related to European ancestry; negatively related to African ancestry, and unrelated to Amerindian ancestry. Notably, the correlations between cognitive ability, parental education, and European and African ancestry are higher than reported by Lasker et al. (2019) for their monoracial African American sample. This is likely due to the greater variance in admixture among Hispanics in this sample (SD_{European} = 29%). The correlation between European ancestry and parental education in this sample was also higher than that reported for European ancestry and socioeconomic status among Puerto Ricans (*r* = .16, *K* = 3, *N* = 1,943; Kirkegaard, Wang, & Fuerst, 2017).

Table S1. Pairwise Correlations among Self-Identified Hispanic-Americans.

	Cognitive Ability	Parental Education	Euro. Ancestry	Afr. Ancestry	Amer. Ancestry	SIRE European	Color	EduPGS
Cognitive Ability	1 (506)							
Parental Education	.287* (500)	1 (507)						
European Ancestry	.300* (506)	.239* (507)	1 (515)					
African Ancestry	-.294* (506)	-.210* (507)	-.874* (515)	1 (515)				

Amerindian Ancestry	.015 (506)	-.035 (507)	-.160* (515)	-.340* (515)	1 (515)			
SIRE: European	.199* (506)	.223* (507)	.643* (515)	-.544* (515)	-.139* (515)	1 (515)		
Color	-.161* (391)	-.077 (392)	-.670* (398)	.637* (398)	-.009 (398)	-.409* (398)	1 (398)	
EduPGS	.293* (506)	.330* (507)	.557* (515)	-.544* (515)	.027 (515)	.355* (515)	-.385* (398)	1 (515)

Note: *Significant at $p < 0.01$. Pairwise *N* in parentheses; EduPGS = MTAG 10k EduPGS.

Table S2 additionally shows the correlations for the combined group. As expected, the correlations for European ancestry are higher in the combined sample, since there is greater variability. The correlation between European ancestry and parental education was also higher than that between European ancestry and SES (*r* = .17, *K* = 15, *N* = 15,980.50; Kirkegaard, Wang, and Fuerst, 2017) previously reported for multi-ethnic and/or unspecified North and Latin American samples. This may again be due to this sample's higher variability in ancestry.

Table S2. Pairwise Correlations among all participants in this sample.

	Cognitive Ability	Parental Education	Euro. Ancestry	Afr. Ancestry	Amer. Ancestry	SIRE European	Color	EduPGS
Cognitive Ability	1 (7,920)							
Parental Education	.401** (7,846)	1 (7,846)						
European Ancestry	.405** (7,920)	.403** (7,921)	1 (8,009)					
African Ancestry	-.406** (7,920)	-.400** (7,921)	-.988** (8,009)	1 (8,009)				
Amerindian Ancestry	-.025* (7,920)	-.047** (7,921)	-.150** (8,009)	-.003 (8,009)	1 (8,009)			
SIRE: European	.396** (7,920)	.405** (7,921)	.938** (8,009)	-.930** (8,009)	-.124** (8,009)	1 (8,009)		

Color	-.350** (5,991)	-.341** (5,989)	-.865** (6,050)	.859** (6,050)	.102** (6,050)	-.806** (6,050)	1 (6,050)	
EduPGS	.399** (7,920)	.442** (7,921)	.667** (8,009)	-.666** (8,009)	-.058** (8,009)	.629** (8,009)	-.603** (6,050)	1 (8,009)

Note: **Values are significant at $p < 0.0001$; *values are significant at $p < 0.05$. Pairwise N in parentheses; EduPGS = MTAG 10k EduPGS.

2. Cognitive Ability, Parental Education, and Education-related PGS (eduPGS) by SIRE Group

We next report bivariate associations between cognitive ability, parental education, and four eduPGS from Lee et al. (2018). Hispanic results appear in Table S3. And the results for European, Europe-African, and African Americans are shown in Tables S4 to S6. In the Hispanic-only sample, all but the putatively causal eduSNPs were significantly associated with cognitive ability. Moreover, there were no statistically significant differences in the magnitudes of the correlations between Hispanic and European Americans for the putative causal eduPGS, the GWAS eduPGS, and the MTAG 10K PGS. However, the MTAG-lead PGS were significantly more predictive for Hispanics ($z = 2.11$, two-tail-$p = 0.0349$). In the case of GWAS eduPGS, the lack of difference in predictivity is somewhat surprising, since this is predicted to show high LD decay, and thus relatively low predictivity in non-European samples. However, it is possible that the association is spuriously high in the Hispanic sample owing to confounding with genetic ancestry. Among both European and European-African Americans, all eduPGS were significantly predictive of *g*. Since the European-African group was 79% European in ancestry, the effect of LD decay may have been minimal. For African Americans, in contrast, the validities were markedly reduced (e.g., MTAG 10k: $r_{\text{European}} = .227$ vs. $r_{\text{African}} = .112$). This is consistent with the general finding of reduced PGS validity among Afro-descent groups (Duncan et al., 2019).

Table S3. Pairwise Correlations Between Cognitive Ability and Education/Intelligence Related Polygenic Scores among Hispanic-Americans.

	Cognitive Ability	**Parental Education**	**Putative Causal**	**GWAS_edu PGS**	**MTAG_10K _eduPGS**	**MTAG_Lead eduPGS**
Cognitive Ability	1 (506)					
Parental Education	.287*** (500)	1 (507)				
Putative Causal_edu PGS	.085 (506)	.084 (507)	1 (515)			
GWAS_edu PGS	.291*** (506)	.281*** (507)	.195*** (515)	1 (515)		
MTAG_10K_ eduPGS	.293*** (506)	.330*** (507)	.255*** (515)	.695*** (515)	1 (515)	
MTAG_Lead_ PGS	.302*** (506)	.334*** (507)	.295*** (515)	.693*** (515)	.868*** (515)	1 (515)

Note: **p* < .05, ***p* <.01, ****p* <.001.

Table S4. Pairwise Correlations Between Cognitive Ability and Education/Intelligence Related Polygenic Scores among European-Americans.

	Cognitive Ability	**Parental Education**	**Putative Causal**	**GWAS_edu PGS**	**MTAG_10K _eduPGS**	**MTAG_Lead eduPGS**
Cognitive Ability	1 (4914)					
Parental Education	.297*** (4886)	1 (4909)				
Putative Causal_edu PGS	.058*** (4914)	.094*** (4909)	1 (4939)			
GWAS_edu PGS	.226*** (4914)	.306*** (4909)	.217*** (4939)	1 (4939)		
MTAG_10K_ eduPGS	.227*** (4914)	.288*** (4909)	.315*** (4939)	.646*** (4939)	1 (4939)	

MTAG_Lead_PGS	.210*** (4914)	.249*** (4909)	.348*** (4939)	.575*** (4939)	.837*** (4939)	1 (4939)

Note: **p* < .05, ***p* <.01, ****p* <.001.

Table S5. Pairwise Correlations Between Cognitive Ability and Education/Intelligence Related Polygenic Scores among European-African Americans.

	Cognitive Ability	**Parental Education**	**Putative Causal**	**GWAS_edu PGS**	**MTAG_10K_eduPGS**	**MTAG_Lead eduPGS**
Cognitive Ability	1 (228)					
Parental Education	.391*** (227)	1 (230)				
Putative Causal_edu PGS	.170* (228)	.201* (230)	1 (232)			
GWAS_edu PGS	.302*** (228)	.400*** (230)	.276*** (232)	1 (232)		
MTAG_10K_eduPGS	.308*** (228)	.381*** (230)	.453*** (232)	.734*** (232)	1 (232)	
MTAG_Lead_PGS	.312*** (228)	.409*** (230)	.462*** (232)	.736*** (232)	.895*** (232)	1 (232)

Note: **p* < .05, ***p* <.01, ****p* <.001.

Table S6. Pairwise Correlations Between Cognitive Ability and Education/Intelligence Related Polygenic Scores among African-Americans.

	Cognitive Ability	**Parental Education**	**Putative Causal**	**GWAS_edu PGS**	**MTAG_10K_eduPGS**	**MTAG_Lead eduPGS**
Cognitive Ability	1 (2179)					
Parental Education	.256*** (2140)	1 (2180)				
Putative Causal_edu PGS	.031 (2179)	.025 (2180)	1 (2228)			

GWAS_edu PGS	.044* (2179)	.032 (2180)	.134*** (2228)	1 (2228)		
MTAG_10K_eduPGS	.112*** (2179)	.119*** (2180)	.227*** (2228)	.482*** (2228)	1 (2228)	
MTAG_Lead_PGS	.095*** (2179)	.117*** (2180)	.266*** (2228)	.451*** (2228)	.800*** (2228)	1 (2228)

Note: **p* < .05, ***p* <.01, ****p* <.001.

References.

Duncan, L., Shen, H., Gelaye, B., Meijsen, J., Ressler, K., Feldman, M., ... & Domingue, B. (2019). Analysis of polygenic risk score usage and performance in diverse human populations. *Nature Communications*, 10(1), 3328.

Kirkegaard, E. O., Wang, M., & Fuerst, J. (2017) Biogeographic Ancestry and Socioeconomic Outcomes in the Americas: a Meta-analysis. *Mankind Quarterly*, 573, 398–427. 9.

Lasker, J., Pesta, B. J., Fuerst, J. G., & Kirkegaard, E. O. (2019). Global Ancestry and Cognitive Ability. *Psych*, 1(1), 431-459.

Lee, J. J., Wedow, R., Okbay, A., Kong, E., Maghzian, O., Zacher, M., ... & Fontana, M. A. (2018). Gene discovery and polygenic prediction from a 1.1-million-person GWAS of educational attainment. *Nature Genetics*, 50(8), 1112.

Via, M., Gignoux, C. R., Roth, L. A., Fejerman, L., Galanter, J., Choudhry, S., ... & Ziv, E. (2011). History shaped the geographic distribution of genomic admixture on the island of Puerto Rico. *PLoS One*, 6(1), e16513.

Supplementary File 4 for "More Research Needed: There is a Robust Causal vs. Confounding Problem for Intelligence-associated Polygenic Scores in Context to Admixed American Populations":

1. Detailed Path Analysis Results

While fitting cross-sectional data to a path model cannot prove the causal assumptions, doing so can provide estimates of the effect magnitudes on the assumption that the model is correct (Bollen & Pearl, 2013). As such, we depict two sets of path model results fit with the **lavaan** R package (Rosseel, 2012). Table S1 shows the path estimates for Hispanics with European ancestry, color, and eduPGS as covariates. Table S2 shows the path estimates for the same model as above but using the complete sample. As an alternative model, Table S3 shows the path estimates for Hispanics with European ancestry, Parental Education, and eduPGS as covariates. Table S4 shows the path estimates for the same model as above but using the complete sample.

Table S1. Detailed Results for the Path Diagram between European Ancestry, Color, eduPGS, and g for Hispanics.

			Unstandardized Estimate	S.E.	*P* value	Lower 95% CI	Upper 95% CI	Standardized Estimate
EUR	→	*g*	0.924	0.260	0.000	0.414	1.433	0.256
EUR	→	eduPGS	2.442	0.168	0.000	2.112	2.773	0.591

eduPGS	→	*g*	0.200	0.051	0.000	0.101	0.299	0.229
Skin Color	→	*g*	0.117	0.075	0.118	-0.030	0.264	0.098
EUR	→	Skin Color	-2.032	0.114	0.000	-2.256	-1.809	-0.670
Skin Color	~	eduPGS	0.009	0.032	0.774	-0.054	0.073	0.015

Note: EUR = European ancestry. Tilde designates covariance.

Table S2. Detailed Results for the Path Diagram between European Ancestry, Color, eduPGS, and g for the Combined Sample.

			Unstandardized Estimate	S.E.	P value	Lower 95% CI	Upper 95% CI	Standardized Estimate
EUR	→	*G*	0.718	0.079	0.000	0.563	0.872	0.232
EUR	→	eduPGS	2.384	0.032	0.000	2.320	2.447	0.688
eduPGS	→	*G*	0.222	0.014	0.000	0.194	0.250	0.248
Skin Color	→	*G*	-0.001	0.026	0.983	-0.051	0.050	0.000
EUR	→	Skin Color	-2.404	0.018	0.000	-2.440	-2.369	-0.865

Skin Color	~	eduPGS	-0.008	0.006	0.176	-0.019	0.004	-0.017

Note: EUR = European ancestry. Tilde designates covariance.

Table S3. Detailed Results for Path Diagram with Parental Education for Hispanics.

			Unstandardized Estimate	S.E.	P value	Lower 95% CI	Upper 95% CI	Standardized Estimate
EUR	→	*G*	0.727	0.193	0.000	0.348	1.105	0.188
EUR	→	eduPGS	2.162	0.144	0.000	1.881	2.443	0.559
eduPGS	→	*G*	0.123	0.051	0.016	0.023	0.224	0.123
Parental Education	→	*G*	0.225	0.049	0.000	0.129	0.320	0.201
EUR	~	Parental Education	0.071	0.013	0.000	0.044	0.097	0.242
Parental Education	~	eduPGS	0.217	0.042	0.000	0.135	0.299	0.232

Note: EUR = European ancestry. Tilde designates covariance.

Table S4. Detailed Results for Path Diagram with Parental Education for the Combined Sample.

			Unstandardized Estimate	S.E.	P value	Lower 95% CI	Upper 95% CI	Standardized Estimate
EUR	→	*G*	0.621	0.041	0	0.539	0.702	0.2
EUR	→	eduPGS	2.149	0.027	0	2.096	2.202	0.666

eduPGS	→	*G*	0.147	0.013	0	0.121	0.172	0.153
Parental Education	→	*G*	0.286	0.012	0	0.261	0.31	0.254
EUR	~	Parental Education	0.145	0.004	0	0.137	0.154	0.401
Parental Education	~	eduPGS	0.203	0.009	0	0.185	0.221	0.232

Note: EUR = European ancestry. Tilde designates covariance.

2. Vectors Vector Correlations for MCV analysis

The Method of Correlated Vectors (MCV) involves correlating two vectors containing subtest related effects (e.g., subtest heritability and subtest *g*-loading). Table S5 shows the subtest correlations used for this analysis. Table S6 shows the SIRE scores, which were used to calculate vectors of standardized group differences.

Table S5. Rounded Correlation Vectors for the Method of Correlated Vector Analysis.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Test	*g* loadings	Anc. *r*	Anc. *r* HI	Anc. *r* AA	PGS *r*	PGS EA *r*	PGS HI *r*	PGS AA *r*	h^2
PLOT*	0.63	0.12	0.25	0.10	0.16	0.13	0.25	0.10	0.30
PCPT*	0.32	0.07	0.06	0.01	0.11	0.09	0.09	0.00	0.33
PCET*	0.45	0.09	0.19	0.01	0.11	0.08	0.20	0.01	0.06

LNB*	0.50	0.08	0.12	0.04	0.12	0.11	0.09	0.02	0.28
VOLT*	0.43	0.09	0.15	0.04	0.10	0.09	0.09	0.04	0.26
TAP	0.33	0.00	0.04	0.04	0.04	0.03	0.14	0.06	0.31
PMAT*	0.63	0.10	0.22	0.05	0.18	0.16	0.24	0.10	0.38
MP	0.27	0.02	0.04	0.01	0.04	0.03	0.05	0.01	0.18
PEDT	0.46	0.04	0.05	0.06	0.05	0.05	-0.02	0.05	0.26
PVRT*	0.71	0.17	0.23	0.07	0.23	0.19	0.24	0.10	0.47
PEIT	0.33	0.00	0.06	0.01	0.00	0.00	0.00	0.04	0.32
PFMT*	0.25	-0.02	0.02	0.01	0.01	0.01	0.02	0.01	0.42
PADT	0.39	0.01	0.03	0.04	0.03	0.04	-0.03	0.02	0.14
PWMT*	0.39	0.04	0.06	0.03	0.07	0.07	-0.04	0.03	0.21
WRAT*	0.63	0.14	0.27	0.06	0.25	0.22	0.28	0.12	0.70

Note: (1) Subtest *g*-loading, (2) correlation between subtest scores and European ancestry (combined sample), (3) correlation between subtest scores and European ancestry (Hispanic sample), (4) correlation between subtest scores and European ancestry (African sample), (5) correlation between subtest scores and eduPGS (combined sample), (6) correlation between subtest scores and eduPGS (European sample), (7) correlation between subtest scores and eduPGS (Hispanic sample), (8) correlation between subtest scores and eduPGS (African sample), average heritability based on the European and African American samples. *Denotes that the subtests were used in computing *g* scores.

Table S6. General Intelligence (g) and Subtest Means and Standard Deviations for European (EA), European-African (EA-AA), African (AA), and Hispanic (HI) Americans.

	EA		EA-AA		AA		HI	
g	0.00	1.01	-0.14	1.05	-1.01	1.07	-0.57	1.13
PLOT*	-0.01	1.00	-0.22	0.96	-0.71	0.96	-0.36	1.04

PCPT*	0.00	1.00	-0.04	1.03	-0.33	1.16	-0.26	1.26
PCET*	0.00	1.00	0.03	0.97	-0.45	1.08	-0.26	1.08
LNB*	0.00	1.00	-0.02	0.98	-0.47	1.18	-0.29	1.12
VOLT*	0.00	1.00	-0.09	0.96	-0.35	1.11	-0.27	1.06
TAP	0.00	1.00	0.04	1.07	-0.07	1.06	0.02	0.98
PMRT*	0.01	1.00	-0.06	1.05	-0.56	0.94	-0.25	0.98
MP	0.00	1.00	-0.16	1.10	-0.11	1.04	-0.07	1.18
PEDT	0.00	1.01	-0.06	0.95	-0.15	1.16	-0.15	1.09
PVRT*	0.00	1.00	-0.15	1.07	-0.98	1.13	-0.58	1.18
PEIT	-0.01	1.01	-0.02	1.07	-0.05	1.10	0.02	1.02
PFMT*	-0.01	1.00	0.10	1.03	-0.04	1.08	0.06	1.04
PADT	0.00	1.01	-0.06	0.93	0.00	1.12	-0.06	1.02
PWMT*	0.00	1.01	0.07	1.03	-0.17	1.23	-0.11	1.16
WRAT*	0.00	1.00	-0.13	1.08	-0.85	0.95	-0.48	1.05

Note: *Denotes that the subtests were used in computing *g* scores.

References.

Bollen K. A., Pearl J. (2013). Eight myths about causality and structural equation models. In: Morgan SL, editor. *Handbook of Causal Analysis for Social Research*. Dordrecht, Neth: Springer; 2013. pp. 301–28

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, 48(2), 1-36. URL http://www.jstatsoft.org/v48/i02/

Supplementary File 5 for "More Research Needed: There is a Robust Causal vs. Confounding Problem for Intelligence-associated Polygenic Scores in Context to Admixed American Populations":

Evaluation of Bias and Validity Using the 1000 Genomes Populations

Since ascertainment bias and confounding related to population stratification is of significant concern, we ran supplementary analyses which leveraged the 1000 Genomes data to explore the effects of score construction on the magnitudes of population differences.

1.1 Population-GWAS vs. Within family Weights

PGS scores are calculated by weighting the trait-associated SNP allele frequencies by each SNP's effect on the predicted trait (i.e., the SNP βs). However, population structure may bias both the βs and SNP selection. This form of bias can be partially circumvented by using βs calculated from within-family analyses, which are robust to the effects of population structure (Sohail et al., 2019). However, using only within-family Betas does not completely address the problem of population stratification, since there could be bias due to SNP selection (Zaidi and Mathieson, 2020). In this analysis, we compare the differences between eduPGS computed with (1) population-GWAS vs. (2) population-GWAS SNPs & within family βs weights vs. (3) within family SNPs & within family βs weights.

To examine the impact of using population-GWAS versus within family Beta weights, we created eduPGS for both CEU and YRI individuals using the Population-GWAS Betas and the within-family Betas. We further decomposed the scores by ancestral and derived status. Next, we extended the population-GWAS versus within family Beta weights to all 1000 Genomes European and African populations. Finally, to address the concern raised by Zaidi and Mathieson (2020), we computed eduPGS based on both within family SNPs and within family βs weights.

1.2 Methods

Lee et al. (2018) report the βs for the 10k MTAG SNPs based on their analysis of 1.1 million (mostly) unrelated individuals. The predicted traits were cognitive ability, self-reported math ability, and highest math class taken. On request, the authors also provided the βs for their analysis of 22,000 sibling pairs. The predicted trait was self-reported years of education. As Lee et al. (2018) note, these within-family estimates are smaller than the corresponding estimates from the population GWAS. The authors explore different reasons for this (Suppl. Note *pp*. 21-38) and reason that the lower validity is likely due in part to a within-family reduction of gene-by-environmental correlation.

We first computed population-GWAS and within family Beta weighted eduPGS for the 1000 Genomes Northern and Central European descent from Utah (CEU) and Yoruba Nigerian (YRI) samples. This was done separately for each individual, so we could get means and standard deviations. We used Europeans and Africans because the ancestral populations of the admixed groups were primarily European and African in origin. Before computing eduPGS, we filtered the 10k MTAG SNPs to those for which both population-GWAS and within family βs were available; thus the population-GWAS and the within family Beta weighted eduPGS use the same set of SNPs. Moreover, the SNPs were filtered for MAF >0.01 for both CEU and YRI. We then repeated this analysis for the 5 unadmixed European and 5 unadmixed African 1000 Genomes populations, using the same method as above. Finally, we computed within family weighted eduPGS based on the 4,413 within family SNPs that had a *p*-value < .05, using the same MAF filter as above.

1.3 Results

Figure S1 and S2 depict, respectively, the population-GWAS and within family weighted eduPGS for CEU and YRI individuals. The difference in betas came to β = 1.66 and β = 1.18, respectively. (Note, these βs are based on total sample *SDs*, which are larger than the average of the within population *SDs*). In both cases, the differences were highly significant and large by conventional interpretative standards. This difference between the two represents a 29% reduction in the eduPGS gap size. For comparison, Lee et al. (2018) report that the within-family effect sizes are

40% smaller than the population-GWAS ones (among Europeans). Thus, this magnitude of reduction in gap size is consistent with the reduction in validity among Europeans.

Figure S1: Plot of population-GWAS Weighted, MTAG SNP eduPGS for CEU and YRI 1000 Genomes Individuals.

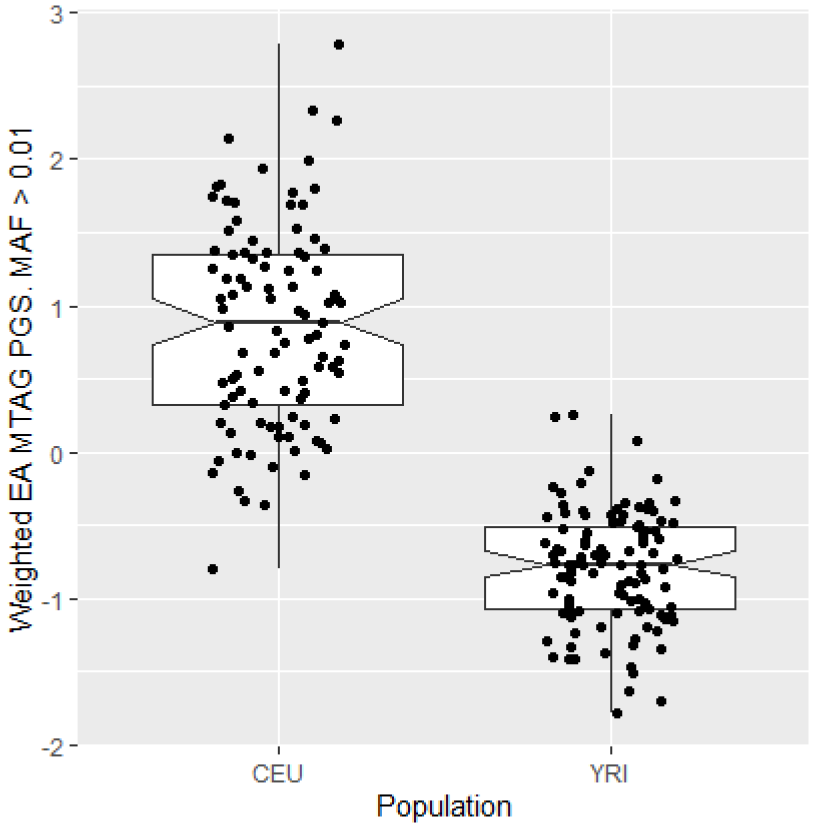
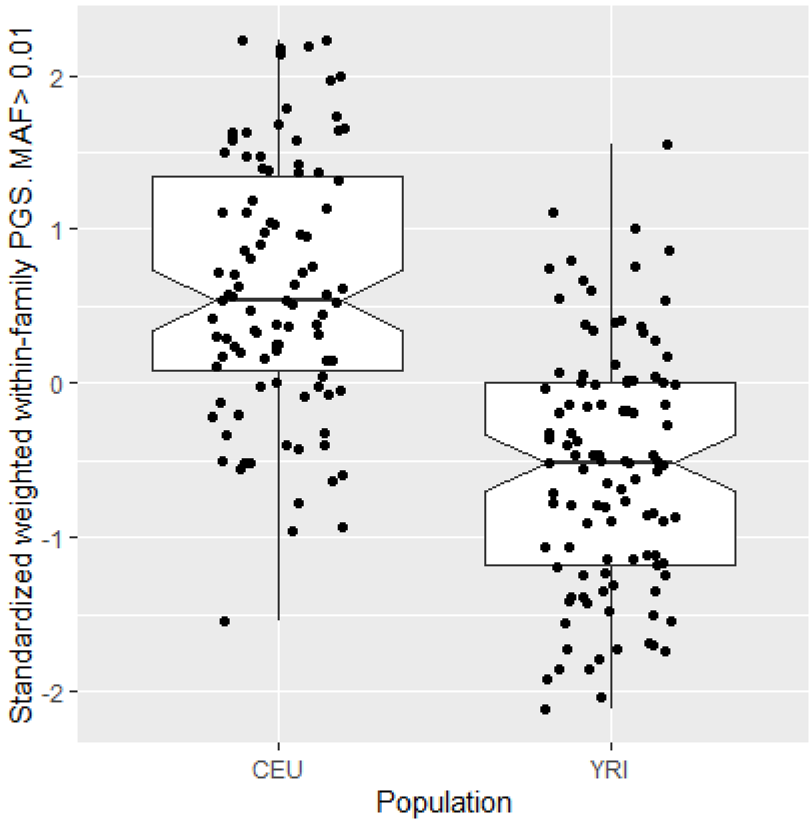


Figure S2: Plot of Within-family, Weighted MTAG SNP eduPGS for CEU and YRI 1000 Genomes Individuals.



Further, Table S1 gives the eduPGS means for all 10K MTAG SNPs, for the derived SNPs, and for the ancestral SNPs. For all sets, the differences are significant and large, as determined using Welch's Two Sample *t*-test.

Table S1. Mean MTAG-based PGS for CEU and YRI Calculated using population-GWAS and Within Family Betas.

	W/ population-GWAS		W/ Within Family Betas	
	CEU (N = 99)	YRI (N = 108)	CEU (N = 99)	YRI (N = 108)
All SNPS	0.866	-0.794	0.614	-0.563
p-value (Welch's Two Sample *t*-test)		< 0.0001		< 0.0001
Derived SNPs	0.938	-0.860	0.702	-0.643
p-value (Welch's Two Sample *t*-test)		< 0.0001		< 0.0001
Ancestral SNPs	0.605	-0.554	0.528	-0.484
p-value (Welch's Two Sample *t*-test)		< 0.0001		< 0.0001

Note: SNPs were filtered for MAF >0.01 for both CEU and YRI. Scores represent standard scores calculated using the standard deviation in the total sample. Sample sizes for the *t*-test were N = 99 for CEU and N=108 for YRI.

We next computed the MTAG-based PGS for the 5 African and 5 European 1000 Genomes populations. These results are shown below in Table S2. As seen in Table S2, there are large eduPGS differences between the 1000 Genomes European and African populations. The European-African difference in betas for population-GWAS and within family weighted MTAG PGS came to β = 1.66 and β = .93, respectively. This difference between the two represents a 44% reduction in the eduPGS gap size. Note, the standardized scores were computed using the total sample, so the scores for CEU and YRI are different in Table S1 (with 2 populations) than in S2 (with 10 populations); also, the standard deviations (within populations) is less than 1.0, because a significant portion of the variance was between populations, so the β difference is not an effects size which uses average or pooled *SDs*.

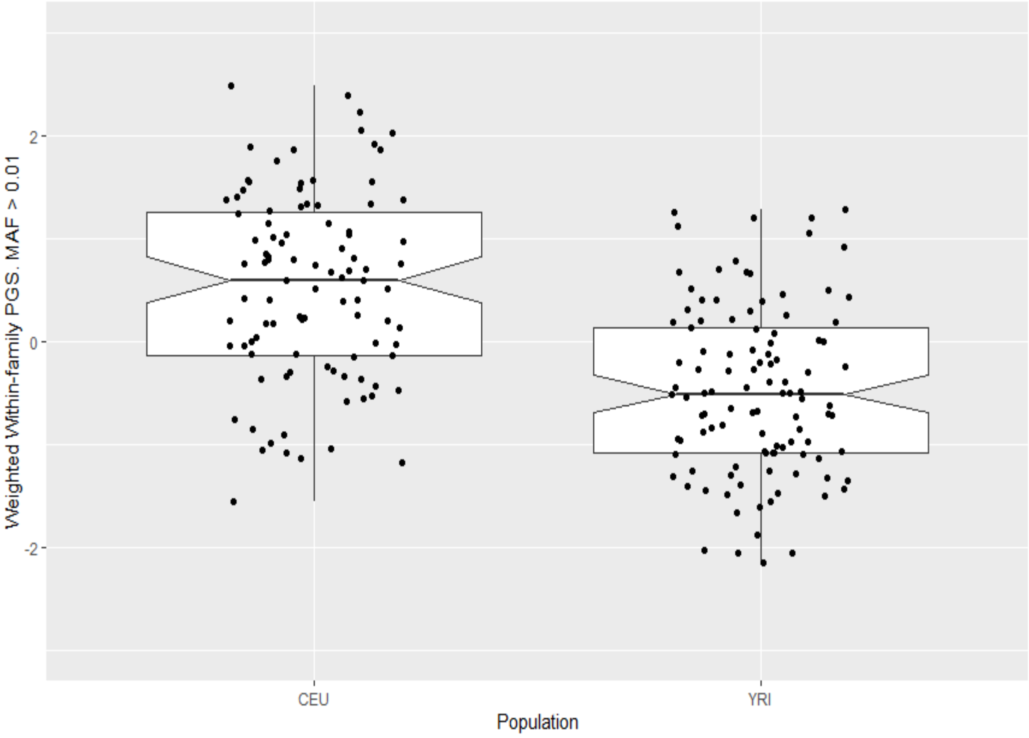
Table S2. Mean MTAG-based PGS for European and African 1000 Genomes populations calculated using population-GWAS and Within Family Betas.

		MTAG SNP Frequencies (Unweighted)		MTAG Standard Scores (population-GWAS Weights)		MTAG Standard Scores (Within Family Weights)	
Population	*N*	*M*	*SD*	*M*	*SD*	*M*	*SD*
CEU	99	0.510	0.0063	0.800	0.6550	0.707	0.7530
FIN	99	0.510	0.0105	0.876	0.8480	0.345	1.0700
GBR	91	0.509	0.0064	0.694	0.6310	0.498	0.8080
IBS	107	0.511	0.0059	0.949	0.6440	0.359	0.7670
TSI	107	0.509	0.0055	0.806	0.5560	0.439	0.8120
EUR_Average	503	0.510	0.0071	0.829	0.6723	0.467	0.8488
YRI	108	0.494	0.0040	-0.768	0.3740	-0.365	0.7240
ESN	99	0.492	0.0043	-0.839	0.4200	-0.405	0.7100
GWD	113	0.493	0.0039	-0.795	0.4010	-0.553	0.6660
LWK	99	0.492	0.0055	-0.839	0.4740	-0.522	1.4900
MSL	85	0.492	0.0040	-0.919	0.3930	-0.484	0.6990
AFR_Average	504	0.493	0.0044	-0.827	0.4133	-0.466	0.9107

Note: SNPs were filtered for MAF >0.01 for both CEU and YRI. Scores represent standard scores calculated using the means and standard deviation in the total sample. Note also, populations *SDs* are less than 1 because these *SDs* include only within-population variance.

Further, to see if use of MTAG SNPs were biasing the results, we computed the differences using the 4,413 within-family SNPs that had a *p*-value < .05 along with the within family weights. These thus are pure within-family based eduPGS and so should show no population structure related bias. The βs for CEU and YRI were 0.529 and -.485, respectively with a β difference of 1.01. Results are shown in Figure S3.

Figure S3: Plot of Within-family Weighted, Within-Family SNP eduPGS for CEU and YRI 1000 Genomes Individuals.



1.4 Interpretation: Population-GWAS vs. Within family EduPGS

Though reduced in size, the within-family weighted eduPGS differences remain large. Moreover, this magnitude of reduction is roughly consistent with the reduced effect sizes that within-family eduPGS have, as compared to population-GWAS ones (among Europeans). Generally, the differences are unlikely to be due to population structure-related bias in the SNP βs. Moreover, they are unlikely to be due completely to population structure-related to SNP selection, since differences were substantial when using both within family weights and SNPs.

2.1 Trans-Ethnically Concordant Betas

Previous polygenic selection studies have applied European GWAS βs to different world populations (e.g., Berg and Coop, 2014). In theory, however, taking into account information about population-specific effects (i.e., βs for both European and non-European comparison samples) should yield more accurate results both on the individual and population levels (Márquez-Luna et al., 2017; Grinde et al., 2018). Indeed, the PGS βs for SNPs in European samples will often show opposite or discordant effects in non-European samples. Since SNPs which show directionally concordant effects across ethnic groups are more likely to be causal (when individual differences are caused by variants common across ethnic groups), computing PGS using SNPs with only concordant effects may either increase or decrease apparent eduPGS gaps.

Given this, it has been argued that including SNPs with transracially discordant effects may bias the group differences and so that "the polygenic scores should be computed only from those GWAS hits that have directionally consistent effects in the races that are being compared" (Thompson, 2019). Thus, for this analysis, we use the two largest TCP samples, European and African Americans, to classify MTAG βs into trans-ethnically concordant and discordant ones. We then recomputed concordant and discordant eduPGS and compared the magnitude of the 1000 Genomes CEU and YRI differences.

2.2 Method

Using the TCP sample, we computed the 10k MTAG SNP betas for *g* separately for European and African Americans. These results are provided in the supplementary excel file. The 10k MTAG SNPs were then split into two sets: 1) concordant SNPs, which had the same direction of effect for TCP African and European Americans and 2) discordant SNPs, which had opposite directions for TCP African and European Americans. Polygenic scores were then computed for the concordant SNPs only and the discordant SNPs only. In computing the eduPGS, the SNPs were weighted with the βs reported by Lee et al. (2018). The rationale was that GWAS βs are more reliable than the TCP βs because they are based on a much larger sample size, thus TCP data is only used to identify concordant / discordant SNPs status. For comparison, eduPGS were additionally computed, with the same weights, based on all 10k MTAG SNPS.

2.3 Results

There were 4,307 and 3,994 concordant and discordant SNPs, respectively. This suggests there is a slight overrepresentation of concordant SNPs. The binomial probability of having 4,307 or more discordant SNPs out of 8,301 is p = 0.0003. Table S3 reports the eduPGS based on all SNPs, the concordant only, and discordant only.

As shown in Table S4, the differences were largest in the trans-ethnically concordant SNPs. In line with the results from 4.1, the discordant PGS showed no CEU-YRI difference (95% C.I. = -0.009, 0.005), while the concordant CEU-YRI difference was around 3% (95% C.I. = 0.025, 0.039). Thus, the presence of discordant variants may possibly be masking CEU-YRI differences as would be the case if individual differences were caused by variants common across ethnic groups and, also, if the transethnic differences were larger for causally-relevant SNPs. However, again, this issue will have to be reevaluated when SNP effect sizes based on larger non-European samples are available. At present, we can only say that given the data available, the eduPGS differences are not inflated as a result of inclusion of SNPs with transethnically discordant effects.

Table S3. Concordant, Discordant, and "Naive" PGS Frequencies by Population.

Population	PGS 10k MTAG	PGS concordant	PGS discordant
CEU	0.5063	0.5050	0.5093
YRI	0.4904	0.4726	0.5109
Difference (CEU - YRI)	0.0159	0.0325	-0.0016

Table S4. Results of t-test for PGS Frequency CEU-YRI difference.

	T	*P*-value	95% CI for Mean Difference	Df
PGS 10k MTAG	6.288	$3.37*10^{-10}$	0.0109, 0.0208	8418
PGS concordant	9.013	$2.2*10^{-16}$	0.0254, 0.0395	4305
PGS discordant	-0.453	0.651	-0.0088, 0.0055	3993

Note: Sample sizes for the t-test were the number of concordant and discordant SNPs. Using the number of individuals, instead, did not change the interpretation of the results.

2.4 Interpretation

The concordant PGS had a higher CEU-YRI difference than both the discordant PGS and the combined eduPGS. In fact, the discordant PGS showed no CEU-YRI difference (95% C.I. = -0.009, 0.005), while the concordant CEU-YRI difference was around 3% (95% C.I. = 0.025, 0.039). A two-way ANOVA showed this interaction to be significant. Generally, the results from this analysis are consistent with the hypothesis that the ability of the MTAG eduPGS to predict population differences in IQ and scholastic achievement is driven by the subset of SNPs with trans-ethnically homogeneous effects and that the discordant SNPs reduce the magnitude of the polygenic score differences. However, it needs to be noted that the power to correctly classify SNPs was low, so this analysis will have to be repeated when better data is available.

3.1. Cross-population Validity

Cross-population analyses are frequently used to assess the validity of PGS (e.g., Berg et al., 2019, Figure 1; Sohail et al., 2019; Figure 4). At times, it is found that PGS differences are directionally inconsistent with observed trait differences across populations. When this is found, the cross-population validity of the PGS is called into question (e.g., Martin et al., 2017). Thus, in the third analysis, we compared the cross-population predictivity for measured population IQ / test scores. In addition to 10K MTAG and MTAG-lead eduPGS, we look at the relation by concordant and discordant status, as it is expected that the discordant 10K MTAG eduPGS will be less predictive than the concordant one, as found on the individual level. Moreover, we examine if eduPGS computed from Lee et al.'s (2018) sibling analysis data predicts national cognitive scores. To do so, we computed eduPGS both using all 81,130 within-family SNPs and the 4,413 within-family SNPs that had a *p*-value < .05. Note, while sibling analyses are robust to population structure-related confounding, none of these SNPs met the minimum for GWAS significance (5e-8) and so the eduPGS computed from them provides a very noisy signal.

3.2 Methods

The Measured population cognitive scores for 18 countries were copied from Lynn and Becker (2019) and World Bank (2017). EduPGS were calculated for the 26 1000 Genomes populations, using the 10K MTAG, MTAG-lead SNPs, the concordant and discordant 10K MTAG SNPs, and within-family based weighted SNP frequencies. The latter were computed using all within-family SNPs along with the within family β weights. Scores for multiple ethnic groups were reported for four countries: USA (European, Mexican, African, and Asian-Indian American), UK (European, Indian, and Sri Lankan British), China (North Han, South Han, and Dai), and Nigeria (Esan and Yoruba). For these, eduPGS were weighted as shown in Table S6 to create national eduPGS. Intra-national group scores were not used as 1) scores are not psychometrically comparable to international ones (Wicherts & Wilhelm, 2007; Täht & Must, 2013), though when available they are reported with the sources noted, and 2) migrant populations can not be assumed to be representative of national ones.

3.3. Results.

Table S5 reports the descriptive statistics.

Table S5: Polygenic and Cognitive Scores at the Country level

Population	MTAG 10k eduPGS	MTAG Lead eduPGS	MTAG Concordant eduPGS	MTAG Discordant eduPGS	Lee et al.'s (2018) Within-	Lee et al.'s (2018) Within-	Ethnic IQs	Lynn & Becker's	World Bank's

						Family (All)	Family (*P* < .05)		(2019) NIQs	(2017) Test Scores
Afr.Car.Barbados		-1.376	-1.343	-1.418	-0.219	0.0000151	0.0002438		91.69/ 90.15*	
Bengali Bangladesh		-0.021	-0.122	0.058	-0.269	0.0000240	0.0003013		74.36	368
Colombian		0.368	0.305	0.557	-0.614	0.0000266	0.0003316		82.99	424
Finland		1.043	0.986	1.117	-0.187	0.0000246	0.0002973		100.55	548
Gambian		-1.469	-1.425	-1.544	-0.115	0.0000141	0.0002370		60	338
Iberian, Spain		0.264	0.186	0.502	-0.562	0.0000262	0.0003050		93.87	514
Japan		0.672	0.993	0.613	0.814	0.0000242	0.0003177		106.43	563
Vietnam		1.024	1.211	0.467	2.137	0.0000234	0.0003104		89.53	519
Luhya, Kenya		-1.513	-1.400	-1.615	-0.058	0.0000145	0.0002505		75.22	455
Mende, Sierra Leone		-1.632	-1.585	-1.672	-0.262	0.0000149	0.0002521		60	316
Peruvian, Lima		-0.500	-0.676	-0.047	-1.540	0.0000286	0.0003246		81.42	407
Punjabi, Pakistan		0.187	0.014	0.284	-0.327	0.0000242	0.0002940		80.05	339
Puerto Rican		0.230	0.190	0.477	-0.925	0.0000251	0.0003124		81.89	
Toscani, Italy		0.907	0.796	1.105	-0.707	0.0000263	0.0003210		94.16	514
Nigeria		-1.471	-1.317	-1.637	0.203	0.0000157	0.0002475		67.83	325
Esan, Nigeria	0.01	-1.545	-1.364	-1.609	-0.167	0.0000142	0.0002343	(??.??)		
Yoruba, Nigeria	0.2	-1.467	-1.315	-1.638	0.222	0.0000158	0.0002481	(??.??)		
USA		0.458	0.324	0.671	-0.824	0.0000268	0.0003099		97.43	523
Utah Whites	0.63	0.924	0.748	1.139	-0.787	0.0000286	0.0003221	(100.00)		
Mexican in L.A.	0.11	-0.077	-0.181	0.293	-1.306	0.0000273	0.0003029	(91.45)		
US Blacks	0.13	-1.347	-1.301	-1.262	-0.664	0.0000178	0.0002575	(85.00)		
Gujarati Indian, Tx	0.01	0.469	0.269	0.443	0.103	0.0000225	0.0003005	(101.65)		
China		1.227	1.496	0.786	2.122	0.0000249	0.0003110		103.95	456
Chinese, Bejing	0.47	1.364	1.631	0.863	2.368	0.0000260	0.0003087	(105.90)		
Chinese, South	0.47	1.089	1.360	0.708	1.876	0.0000238	0.0003134	(105.90)		
Chinese Dai		0.739	1.047	0.491	1.271	0.0000231	0.0003004	(93.90)		
UK		0.783	0.600	1.044	-0.960	0.0000267	0.0003086		99.22	517
British, GB	0.82	0.790	0.609	1.060	-0.993	0.0000268	0.0003090	(100.00)		
Indian Telegu, UK	0.02	0.501	0.249	0.400	0.379	0.0000243	0.0002925	(??.??)		
Sri Lankan, UK		0.376	0.118	0.227	0.531	0.0000250	0.0003024	(??.??)		

Note: USA ethnic scores from Fuerst (2014), with second-generation "Hispanic" substituted for Mexican; Chinese ethnic scores from Lynn and Cheng (2014). USA eduPGS calculated as the weighted eduPGS of European, Mexican, African, and Asian-Indian Americans. Chinese eduPGS calculated as the weighted. eduPGS of North and South Han. UK eduPGS calculated as the weighted eduPGS of White and Indian eduPGS.Nigerian eduPGS calculated as the weighted eduPGS of Esan and Yoruba eduPGS (??.??) indicates unknown intranational scores. *Lynn and Becker (2019) report an IQ of 91.69 for Barbados. However, this becomes 90.15 when including the 114 generation 2 sample WASI scores reported by Wabler et al. (2018). We use this later score.

Table S6 reports the correlation matrix. As shown, the Lead SNP eduPGS has high predictive validity for both Lynn and Becker's National IQs (*r* = .817) and World Bank's Test Scores (*r* = .747). These values were equivalently high for the MTAG 10K eduPGS, at *r* = .804 and *r* = .734, respectively. As predicted, the correlations for the discordant eduPGS, unlike the concordant ones, were low at *r* = .213 (Lynn & Becker, 2019) and *r* = .142 (World Bank, 2017), respectively. The difference between the concordant (*r* = .784) and discordant (*r*

= .213) correlations with Lynn and Becker's NationalIQs was significant (*t* =5.93, *p* <0.01; dependent samples). The within-family based eduPGS, based on all SNPS, also had lower validity for both Lynn and Becker's National IQs (*r* = .648) and World Bank's Test Scores (*r* = .565), as did the ones that had a *p* < .05, with *r* = .638 (Lynn & Becker, 2019) and *r* = .596 (World Bank, 2017), respectively.

Figure S6. Correlation matrix for eduPGS and Population IQs.

	1	2	3	4	5	6	7
1. MTAG 10K eduPGS	1.00						
2. MTAG Lead eduPGS	.988 (18)	1.00					
3. MTAG Concordant eduPGS	.971 (18)	.930 (18)	1.00				
4. MTAG Discordant eduPGS	.246 (18)	.371 (18)	.011 (18)	1.00			
5. Lee et al. (2018) Within Family	.840 (18)	.776 (18)	.921 (18)	-.190 (18)	1.00		
6. Lee et al. (2018) Within Family (*p* < .05)	.853 (18)	.814 (18)	.905 (18)	-.063 (18)	.964 (18)	1.00	
7. Lynn & Becker (2019) NIQ	.804 (18)	.817 (18)	.784 (18)	.213 (18)	.648 (18)	.638 (18)	1.00
8. World Bank (2017) Test Scores	.734 (16)	.747 (16)	.727 (16)	.142 (16)	.565 (16)	.596 (16)	.885 (16)

Note: Sample sizes in parentheses.

Figure S3 shows the regression plot for NIQ and MTAG 10K eduPGS, while figure S4 shows the regression plot for NIQ and the within-family based eduPGS. As seen, the cross population validity of the within-family eduPGS is tenuous. However, this eduPGS may not be reliable as the within-family SNP with the lowest *p* value (1.877e-05) did not even meet the conventional minimum for GWAS significance (5e-8).

Figure S3. Regression Plot for Lynn and Becker's (2019) NIQ and MTAG 10K eduPGS Scores Based on 1000 Genomes Samples.

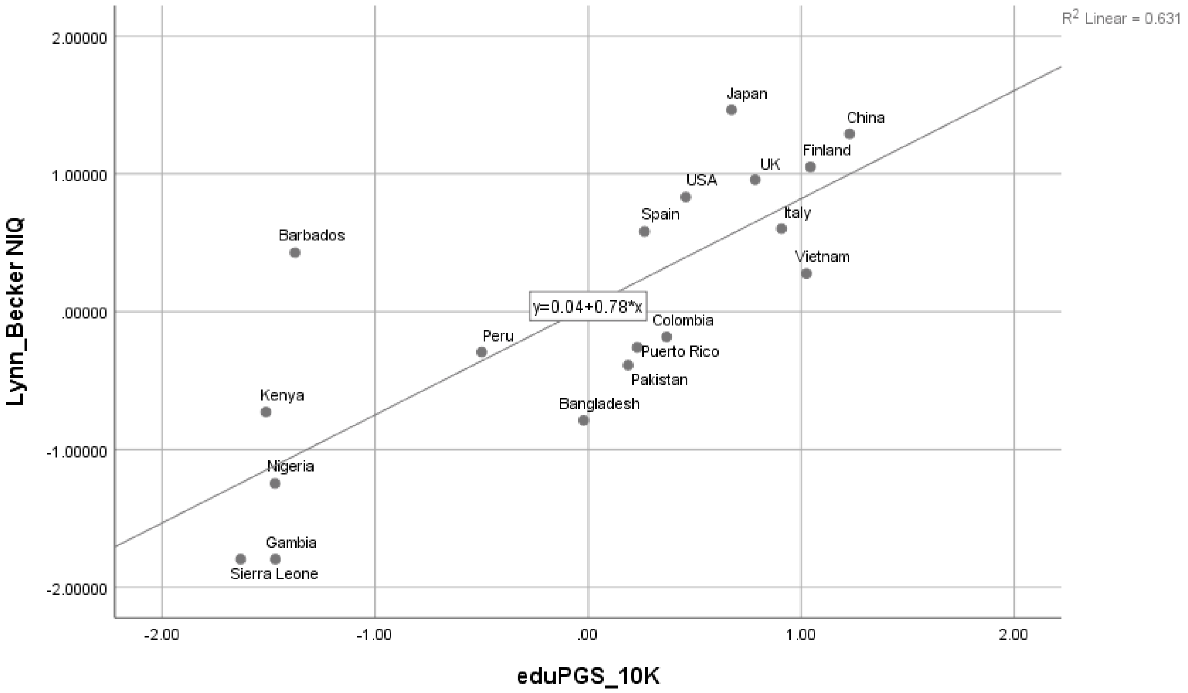
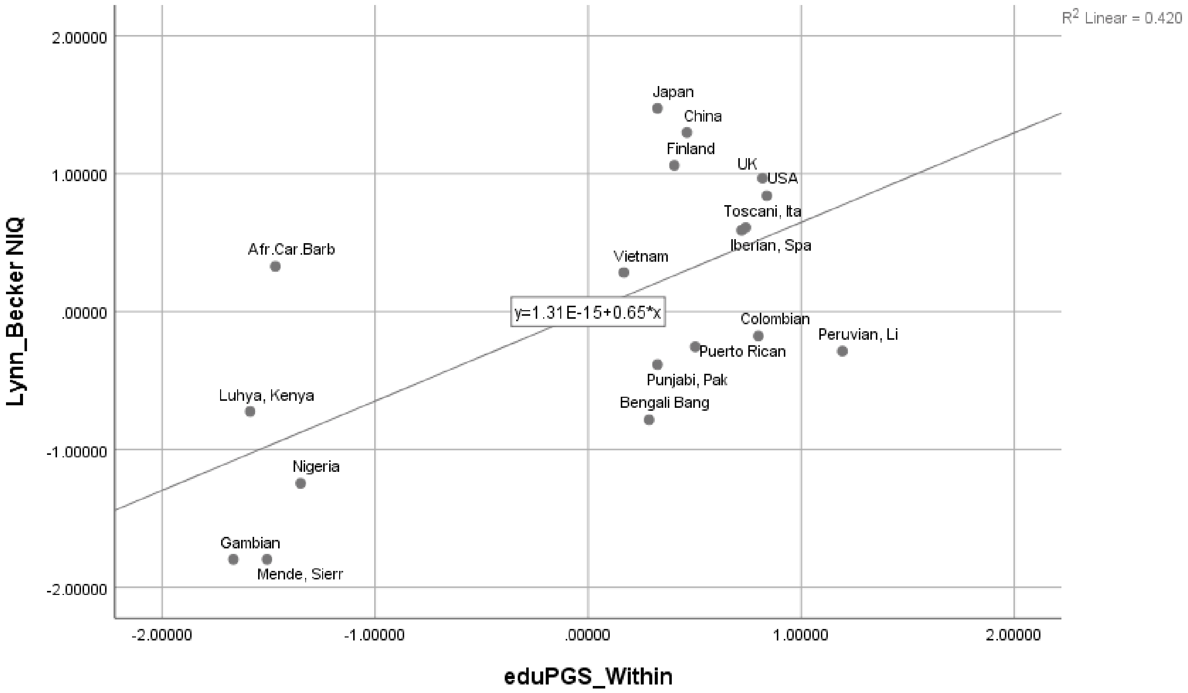


Figure S4. Regression Plot for Lynn and Becker's (2019) NIQ and Within-family Based eduPGS Scores (Using all 81,130 SNPs) Based on 1000 Genomes Samples.



3.4 Interpretation

MTAG-based eduPGS correlates highly with measured cognitive scores. Moreover, the population-level correlations were significantly higher for the concordant than discordant SNPs, consistent with the individual level results. However, the cross-population predictive validity for the within-family based eduPGS, calculated based on 22,000 sibling pairs, is tenuous, with high eduPGS for some low cognitive test scoring countries (e.g., Peru and Colombia). Since none of the SNPs for this eduPGS met the minimum for GWAS significance this may simply be a result of unreliability in the measure. In general, the cross-population validity of the MTAG-based eduPGS can not be rejected on the grounds that eduPGS differences are inconsistent with known phenotypic score differences (e.g., Martin et al., 2017). However, that eduPGS constructed using the smaller within-family sample shows a tenuous validity highlights Duncan et al.'s (2019) caution that different eduPGS can give markedly different results.

References

Berg, J. J., & Coop, G. (2014). A population genetic signal of polygenic adaptation. *PLoS Genetics*, 10(8).

Berg, J. J., Harpak, A., Sinnott-Armstrong, N., Joergensen, A. M., Mostafavi, H., Field, Y., ... & Coop, G. (2019). Reduced signal for polygenic adaptation of height in UK Biobank. *Elife*, 8, e39725.

Duncan, L., Shen, H., Gelaye, B., Meijsen, J., Ressler, K., Feldman, M., ... & Domingue, B. (2019). Analysis of polygenic risk score usage and performance in diverse human populations. *Nature Communications*, 10(1), 1-9.

Fuerst, J. (2014). Ethnic/race differences in aptitude by generation in the United States: An exploratory meta-analysis. *Open Differential Psychology*.

Grinde, K. E., Qi, Q., Thornton, T. A., Liu, S., Shadyab, A. H., Chan, K. H. K., ... & Sofer, T. (2019). Generalizing polygenic risk scores from Europeans to Hispanics/Latinos. *Genetic epidemiology*, *43*(1), 50-62.

Lee, J.J.; Wedow, R.; Okbay, A.; Kong, E.; Maghzian, O.; Zacher, M.; Nguyen-Viet, T.A.; Bowers, P.; Sidorenko, J.; Karlsson Linnér, R.; et al. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nat. Genet.* 2018, 50, 1112–1121.

Lynn, R., & Becker, D. (2019*). The intelligence of nations*. London: Ulster Institute for Social Research.

Lynn, R., & Cheng, H. (2014). Differences in intelligence between ethnic minorities and Han in China. *Intelligence,* 46, 228-234.

Márquez-Luna C, Loh P-R; South Asian Type 2 Diabetes (SAT2D) Consortium; SIGMA Type 2 Diabetes Consortium, Price AL (2017): Multiethnic polygenic risk scores improve risk prediction in diverse populations. Genet Epidemiol 41:811–823.

Sohail, M., Maier, R. M., Ganna, A., Bloemendal, A., Martin, A. R., Turchin, M. C., ... & Neale, B. (2019). Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. *Elife*, *8*, e39702. Täht, K., & Must, O. (2013). Comparability of educational achievement and learning attitudes across nations. *Educational Research and Evaluation*, 19(1), 19-38.

Waber, D. P., Bryce, C. P., Girard, J. M., Fischer, L. K., Fitzmaurice, G. M., & Galler, J. R. (2018). Parental history of moderate to severe infantile malnutrition is associated with cognitive deficits in their adult offspring. *Nutritional Neuroscience*, 21(3), 195-201.

Wicherts, J. M., & Wilhelm, O. (2007). What is the national g-factor. *European Journal of Personality*, 21(5), 763-765.

World Bank (2017). Human Capital Index. Accessed at: https://www.worldbank.org/en/publication/human-capital

Zaidi, A.A. & Mathieson, I.A. (2020). Demographic history impacts stratification in polygenic scores. bioRxiv.

Supplementary File 6 for "More Research Needed: There is a Robust Causal vs. Confounding Problem for Intelligence-associated Polygenic Scores in Context to Admixed American Populations":

Calculating Expected Phenotypic Differences.

1. Formulas

The formal relation between the combined or total within and between group heritability, heritability between groups, and genetic and phenotypic differences is given by Defries (1972a), McClearn and Defries (1973), Loehlin, Lindzey, & Spuhler (1975), Cheverud (1985), Jensen (1998):

$$h^2{}_{\mathrm{G}} = h^2 * \frac{r}{t} \qquad (1)$$

where $h^2{}_{\mathrm{G}}$ is the between group heritability, h^2 is the combined or total heritability, *r* is the genetic intraclass correlation, and *t* is the phenotypic intraclass correlation, which is equivalent to the square of the point biserial correlation (i.e., $r_{\mathrm{pbs}}{}^2$). This formula can be expressed in terms of within groups heritability, $h^2{}_{\mathrm{w}}$. In this case:

$$h^2{}_{\mathrm{G}} = h^2{}_{\mathrm{w}} * \frac{(1-t)r}{(1-r)t} \qquad (2)$$

where $h^2{}_{\mathrm{w}}$ is the average of the heritabilities within both groups. Equations (1) & (2) are simplified, but can be expanded to include gene-environment covariance ($\mathrm{COV_{GE}}$) (Defries, 1972b). In this case, the between group heritability is not $h^2{}_{\mathrm{G}}$ but is equal to:

$$h^2{}_{\mathrm{G}} + h_{\mathrm{G}} * e_{\mathrm{G}} * r\mathrm{A_GE_G}$$

where $\mathrm{h_G}$ and $\mathrm{e_G}$ are the square root of the between groups heritability and between group environmentality, respectively, and $r\mathrm{A_GE_G}$ is the gene-environment correlation between groups. Thus, in the case of positive $\mathrm{COV_{GE}}$, equations (1) and (2) will underestimate genetic differences between groups (McClearn and Defries, 1973). This formula can be further expanded to include dominance (e.g., Wright, 1952). See the exchange between Defries and Jensen (Jensen, 1972; also: Jensen, 1998) for when narrow or broad-sense within groups heritability is more appropriate. Here we will work with the simplified equation.

The intraclass correlations (*r* and *t*) can be interpreted in terms of one-way analysis of variance (Loehlin, Lindzey, & Spuhler, 1975), where:

$$\mathrm{ICC} = \frac{MS\mathrm{b} - MS\mathrm{w}}{MSb + (n-1)\mathrm{MS}w} \qquad (3)$$

where MSb represents the mean square between groups and MSw represents the mean square within groups. ICCs are equivalent to η^2, which can be converted into Cohen’s d with the following equation:

$$\eta^2 = \frac{(.5d)^2}{1 + (.5d)^2} \quad \text{or, equivalently,} \quad d = 2\sqrt{\frac{\eta^2}{1 - \eta^2}} \qquad (4)$$

where Cohen’s *d* is:

$$d = \frac{M1 - M2}{SD_{pooled}} \qquad (5)$$

and M_1 and M_2 are the means for group 1 and group 2, respectively and SD_{pooled} is the pooled standard deviation. Alternatively, η^2 can be converted into a point-biserial correlation, and this can be converted into Cohen’s *d* with the following equation:

$$r_{pbs} = \frac{d}{\sqrt{(d^2 + 4)}} \quad \text{or, equivalently,} \quad d = \frac{2r}{\sqrt{(1 - r^2)}} \qquad (6)$$

For diploid populations, *r*, the genetic intraclass correlation, in equation (1) and (2), is:

$$r = 2F_{st} / (1 + F_{IT}) \qquad (7)$$

where F_{st} is the fixation index, or the between group variance in allele frequencies, and F_{IT} is the overall level of inbreeding in the total population (Hamilton, 1971; Cheverud, 1985; Whitlock, 2004).

Contrary to what is often thought, the upper bounds of F_{st} is typically < 1 (Alcala & Rosenberg, 2017; Alcala & Rosenberg, 2019). Importantly, F_{st} is mathematically constrained by the frequency *M* of the most frequent allele and thus total genetic variance. Thus, Fst is not on the variance scale of 0 to 1 and so is not analogous to ICCs which represent the total portion of variance out of 1. In light of this, Alcala & Rosenberg (2017) proposed the ratio F_{st}/F_{st}_max, which ranges from 0 to 1. To correct for the mathematical constraints and place F_{st} on a variance metric, we can follow (Alcala & Rosenberg, 2017) and create a corrected value, F_{st_c}:

$$F_{st\_c} = F_{st} / F_{st}\_max \qquad (8)$$

The corresponding corrected genetic intraclass correlation, rc , is:

$$r_c = 2 F_{st\_c} / (1 + F_{IT}) \qquad (9)$$

Based on the equations for *M* provided by Alcala & Rosenberg (2017), the F_{st_c} ~ .7, in Table 10 for the pairwise comparisons. However, the values will depend on the subset of populations and SNPs used. There are other concerns with common F_{st} estimators, given assumptions about population structure (Ochoa & Storey, 2021). These assumptions can lead to underestimations of the coefficient of relatedness, which *r* represents (DeFries, 1972a), in context to admixture (Ochoa & Storey, 2019). However, we will proceed with Weir and Cockerham's estimator.

Now, equation (2) can be rearranged to solve for *t* (the phenotypic variance). This gives:

$$t \;=\; h^2{}_{\mathrm{w}} \;\;*\;\; \frac{r}{h^2{}_{\mathrm{G}} - rh^2{}_{\mathrm{G}} + rh^2{}_{\mathrm{w}}} \qquad (10)$$

Based on equation (10), one can solve for the expected gap, where environments are equal, which is done by setting $h^2{}_{\mathrm{G}}$ to 1. This gives the following:

$$t_{\mathrm{expected}} \;=\; h^2{}_{\mathrm{w}} \;\;*\;\; \frac{r}{1 - r + rh^2{}_{\mathrm{w}}} \qquad (11)$$

Using equation (1), with the total heritability instead of within groups heritability, (11) is simply:

$$t_{\mathrm{expected}} \;=\; h^2 * \; r \qquad (12)$$

Equation (12) can be related to the equation for expected differences given by Turkheimer (1991, eq. 6), where:

$$\mathrm{P}_{1\mathrm{observed}} = \sqrt{h^2} \quad \hat{\mathrm{H}}_1 + \sqrt{e^2} \quad \hat{\mathrm{E}}_1 \quad \text{and} \quad \mathrm{P}_{2\mathrm{observed}} = \sqrt{h^2} \quad \hat{\mathrm{H}}_2 + \sqrt{e^2} \quad \hat{\mathrm{E}}_2 \qquad (13)$$

and P_1 and P_2 are the standardized observed phenotypic values for group 1 and group 2, respectively and Ĥ and Ê are the standardized genetic and environmental values for the respective groups.

When $\hat{\mathrm{E}}_1 \;=\; \hat{\mathrm{E}}_2$ (and thus $h^2{}_{\mathrm{G}}$ = 1), then:

$$\mathrm{P}_1 - \mathrm{P}_2 = \sqrt{h^2{}_{\mathrm{w}}}\;(\hat{\mathrm{H}}_1 - \hat{\mathrm{H}}_1)$$

And so, in terms of standardized phenotypic ($\mathrm{d}_{\mathrm{p\ expected}}$) and genetic ($\mathrm{d}_{\mathrm{g}}$) differences:

$$d_{P\ expected} \;=\sqrt{h^2{}_{\mathrm{w}}}\; * dg \qquad (14)$$

With formula (6), we can convert the standardized differences (d_{p} and d_{g}) into point-biserial correlations, yielding:

$$r_{\mathrm{pbs\_phenotypic\ expected}} \;=\; \sqrt{h}^2 \;*\; r_{\mathrm{pbs\_genetic}} \qquad (15)$$

Squaring both sides, recaptures equation (12), since the genetic intraclass correlation (*r*) and the phenotypic intraclass correlation (*t)* are equivalent to the square of the point biserial correlation. From the above, it can also be seen that the $d_{\mathrm{p\ expected}}$ or the "genotypic gap" is equal to $\sqrt{\ } h^2{}_{\mathrm{G}}$ * $\mathrm{d}_{\mathrm{observed,}}$ where the $\sqrt{\ } h^2{}_{\mathrm{G}}$ can be interpreted as the correlation between phenotype and genotype between groups, i.e.:

$$d_{\text{expected}} = \sqrt{h^2{}_{\text{G}}} * d_{\text{observed}} \quad (16)$$

This is because we can rewrite equation (1) as:

$$h^2{}_{\text{G}} * t = h^2 * r \quad (17)$$

Taking the square root of both sides, gives:

$$\sqrt{h}^2{}_{\text{G}} * r_{\text{pbs\_phenotypic}} = \sqrt{h}^2 * r_{\text{pbs\_genetic}} \quad (18)$$

And from equation (15), we see that the left hand is equal to $r_{\text{pbs\_phenotypic expected}}$.

Equation 1 and 2 can be rewritten to solve for $e^2{}_{\text{G}}$, the between group environmentality. This is just 1- $h^2{}_{\text{G}}$, thus:

$$e^2{}_{\text{G}} = 1 - h^2{}_{\text{G}} = 1 - h^2 * \frac{r}{t} \quad (19)$$

To note, while, $e^2{}_{\text{G}}$ and $h^2{}_{\text{G}}$ sum to 1, the expected differences on account of genes and environment, when expressed in standard deviations, will not sum to the phenotypic gap. This is because standard deviations are a linear measurement, and do not express differences in variance units (Jensen, 1998). Rather, to add the effects, one has to take the square root of the sum of the squared differences. The formula is:

$$d_{\text{ phenotypic}} = \sqrt{(d\_\text{genetic})\text{^}2 + (d\_\text{environmental})\text{^}2} \quad (20)$$

For example, in a case where the phenotypic differences is 1 (t = .20), the within groups heritability is .50, and r = .20, $e^2{}_{\text{G}}$ = .5 and $h^2{}_{\text{G}} = .5$. By equation (14), the effect owing to environment will be $\sqrt{.5} * 15$ = .71 SD or 10.6066 IQ points. And the effect owing to genes will be the same; this is also the expected difference given by equation (9). The phenotypic difference is recovered with equation (19), as $\sqrt{10.6066\text{^}2 + 10.6066\text{^}2}$ = 15.

From the above, it is obvious that $h^2{}_{\text{G}}$ is not equal to the real-world percentage of the differences which, owing to genes, would remain when the environments were equalized. The inference makes the R2 interpretative fallacy (Hunter & Schmidt, 2004), which results because variance-explained does not represent a linear relation between x and y. Rather expected percentage genetic, in the ordinary sense, is given by:

$$\text{Percentage genetic expected} = d_{\text{expected}} / \text{d}_{\text{observed}} \quad (21)$$

Example:

Using the Education SNP F_{st} values in Table 1, calculate the expected differences owing to genes for Africans and Europeans, assuming within groups heritabilities of .20 to .80.

Table 1. Fst Values for the 10k MTAG eduSNPs by 1000 Genomes Population Pairs.

Population_1	Population_2	Edu_Fst	Edu_Fit
AFR	EAS	0.1402	0.1470
AFR	EUR	0.1090	0.1153
AFR	SAS	0.1018	0.1125
AFR	AMR	0.0984	0.1160
EAS	EUR	0.0964	0.1030
AMR	EAS	0.0714	0.0899
EAS	SAS	0.0626	0.0741
EUR	SAS	0.0342	0.0451
AMR	SAS	0.0296	0.0528
AMR	EUR	0.0226	0.0412

Note: AFR = African, EAS = East Asian, SAS = South Asian, Eur = European, and AMR = admixed American (Mexican, Puerto Ricans, Colombian, and Peruvian) populations.

For Africans (AFR) and European (EUR), the MTAG SNPS the Fst = .1090. By equation (7), *r* = 2(.1090) /(1+.1153) = .1955. Given a $h^2{}_w$ = .5, then $t_{expected}$ from equation (9) is:

$$t_{expected} = .5 * \frac{.1955}{(.1955)(.5) + 1 - (.1955)} = .1083$$

Given equations (3) and (4), this equals *d* = 0.70 or a 10.46 point difference on a metric with a standard deviation of 15.

Heritability estimates are population specific. For example, in there meta-analysis, Polderman et al. (2015), Table 2, give twin correlations by age. Using Falconer's formula, these convert into H^2s of .46 at 0 to 11 years of age and 0.80 at 18 to 64 years. Since estimates of h^2 and H^2 are population specific, since the specific genetic variance components (e.g., additive, GE covariance, dominance) are often not known, and since there may be disagreements on how to correct F_{st} , one can provide a table for the different possibilities, given F_{st} = .1090. This is shown in Table 2.

Table 2. BGH and Expected Variance and IQ point difference Given Different Values of *r* and $H^{2.}$

H^2	F_{st}	*r*	*t*_observed	BGH	*t*_expected	*d*_expected	Expected IQ point difference	Cohen's Interpretation
0.20	0.1090	0.1955	0.2000	0.194	0.0463	0.4409	6.61	Medium
0.35	0.1090	0.1955	0.2000	0.340	0.0784	0.5833	8.75	Medium
0.50	0.1090	0.1955	0.2000	0.486	0.1083	0.6971	10.46	Medium
0.65	0.1090	0.1955	0.2000	0.632	0.1364	0.7949	11.92	Large
0.80	0.1090	0.1955	0.2000	0.778	0.1628	0.8818	13.23	Large

H^2	F_{st}	*r_c*	*t*_observed	BGH	*t*_expected	*d*_expected	Expected IQ point difference	Cohen's Interpretation
0.20	0.1090	0.2792	0.2000	0.310	0.0719	0.5567	8.35	Medium
0.35	0.1090	0.2792	0.2000	0.542	0.1194	0.7364	11.05	Medium
0.50	0.1090	0.2792	0.2000	0.775	0.1623	0.8802	13.20	Medium
0.65	0.1090	0.2792	0.2000	1.007	0.2011	1.0035	15.05	Large
0.80	0.1090	0.2792	0.2000	1.240	0.2366	1.1133	16.70	Large

Note: see text for definition of variables.

References

Alcala, N., & Rosenberg, N. A. (2017). Mathematical constraints on F ST: biallelic markers in arbitrarily many populations. *Genetics*, 206(3), 1581-1600.

Alcala, N., & Rosenberg, N. A. (2019). , Jost's D, and FST are similarly constrained by allele frequencies: A mathematical, simulation, and empirical study. *Molecular ecology*, 28(7), 1624-1636.

Cheverud, J. M. (1985). A quantitative genetic model of altruistic selection. *Behavioral Ecology and Sociobiology*, 16(3), 239-243.

DeFries, J. C. (1972a). Quantitative aspects of genetics and environment in the determination of behavior. In L. Ehrman, G. Omenn, & E. Caspari (Eds.), *Genetics, Environment, and Behavior* (pp. 6-16). San Diego, CA: Academic Press.

DeFries, J. C. (1972b). Reply to Professor Fuller: J. C. DeFries. In L. Ehrman, G. Omenn, & E. Caspari (Eds.), *Genetics, Environment, and Behavior* (pp. 21-22). San Diego, CA: Academic Press.

Hamilton, W. D. (1971). Selection of selfish and altruistic behavior in some extreme models. *Man and Beast: Comparative Social Bahavior*, 57-91.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. Newbury Park, CA: Sage.

Jensen, A. R. (1972). Comment. In L. Ehrman, G. Omenn, & E. Caspari (Eds.), *Genetics, Environment, and Behavior* (pp. 23-25). San Diego, CA: Academic Press.

Jensen, A. R. (1998). *The g factor*. Westport, CT: Prager.

Loehlin, J.C., Lindzey, G, & Spuhler, J. N. (1975). *Race Differences in Intelligence*. San Francisco: W. H. Freeman.

McClearn, G. E., and De Fries, J. C. (1973). *Introduction to Behavioral Genetics*, Freeman, San Francisco.

Ochoa, A., & Storey, J. D. (2021). Estimating FST and kinship for arbitrary population structures. *PLoS genetics*, 17(1), e1009241.

Ochoa, A., & Storey, J. D. (2019). New kinship and FST estimates reveal higher levels of differentiation in the global human population. *BioRxiv*, 653279.

Polderman, T. J., Benyamin, B., De Leeuw, C. A., Sullivan, P. F., Van Bochoven, A., Visscher, P. M., & Posthuma, D. (2015). Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nature genetics*, 47(7), 702-709.

Turkheimer, E. (1991). Individual and group differences in adoption studies of IQ. *Psychological Bulletin*, 110(3), 392.

Whitlock, M. C. (2004). Selection and drift in metapopulations. *Ecology, Genetics and Evolution of Metapopulations*, 153-173.

Wright, S. (1952). The theoretical variance within and among subdivisions of a population that is in a steady state. *Genetics*, 37(3), 312.

Xu, S., Huang, W., Qian, J., & Jin, L. (2008). Analysis of genomic admixture in Uyghur and its implication in mapping strategy. *The American Journal of Human Genetics*, 82(4), 883-894.