Qeios

Correlations between Socioeconomic Status (SES) and Biogeographic Ancestries: Indirect Evidence of SES Model Misspecification

Gregory Connor[1], John G.R. Fuerst[2] and Meng Hu[1]

[1] Independent researcher

[2] University of Maryland, Department of Biotechnology

Abstract

New genomic technologies allow the apportionment of individuals' genotyped DNA into admixture proportions traceable to historically isolated biogeographic ancestry (BGA) groups such as African, European, and Amerindian. These BGA admixture proportions have proven valuable in a wide range of recent epidemiological research. This paper performs a meta-analysis of these epidemiological studies and finds that, as an ancillary result, these studies reveal consistent patterns of correlation between BGA admixture proportions and socio-economic status (SES). Given this finding, the absence of BGA admixture proportions data from almost all extant economic analysis of individuals' susceptibility to high/low socioeconomic status is indirect evidence for a non-negligible omitted-variable bias in such analysis. Economic models of SES which do not consider BGA as a possible explanatory variable may be unreliable due to the potential confounding associated with this omitted variable.

I. Introduction.

With the completion of the Human Genome Project and subsequent advances in genetic research it is now possible to apportion individuals' genetic ancestry into admixture proportions traceable to historically isolated biogeographic ancestry (BGA) groups such as African, European, and Amerindian. These BGA admixture proportions are a powerful research tool in epidemiology; in a regression model with a health trait as dependent variable and with admixture proportions as explanatory variables, the coefficients on the admixture proportions provide a useful measure of BGA-related genetic variation associated with the health trait. In addition to admixture proportions, it is important to include other health-relevant explanatory variables such as socio-economic status (SES) in the admixture regression. Admixture regression has been used to study alcohol dependence (Zuo et al., 2009), height (Becker et al., 2011), asthma risk (Flores et al., 2012), cardiovascular disease (Bidulescu et al., 2014), sleep depth (Halder et al., 2015),

https://doi.org/10.32388/5OULH0

cigarette smoking behavior (Choquet et al., 2021), metabolomics (Mehanna et al., 2022), cancer (Rhead et al., 2022), and diabetes (Parcha et al., 2023).

Since BGA admixture proportions and SES are both important explanatory variables in epidemiological studies using admixture regression, it is standard statistical practice in such studies to estimate the correlation between them, as a monitoring device on regression model misspecification or under-identification. These correlation estimates linking BGA admixture proportions and SES are only a tangential concern within epidemiology but have considerable interest for other research fields. As shown below, meta-analysis shows consistent patterns in these estimated correlations across studies. These correlation patterns have particular relevance to economic models concerned with explaining SES outcomes across individuals; BGA admixture proportions are a potential omitted variable in such models.

This paper empirically explores how BGA admixture proportions and SES are systematically related via a meta-analysis of extant epidemiological studies from the Americas which include estimates of BGA-SES associations. European BGA admixture proportions showed a positive correlation with SES indicators, $r = .16$ (95% confidence interval: .13 to .19), whereas both Amerindian and African BGA admixture proportions were negatively correlated at $r = -.11$ (95% confidence interval: -.15 to -.06) and $r = -.13$ (95% confidence interval: -.17 to -.09), respectively. The same pattern emerges in examining the sign of the estimated association (correlation or other non-correlation statistics such as ANOVA or odds ratios giving directional association) across study samples: 58 out of 68 (85%) of the European BGA-SES estimated associations are positive, 2 of the 68 (3%) are negative, and in the remaining 8 of the 68 samples (12%) the results are indeterminate with no clear direction across measures of association within the study. In the case of Amerindian BGA-SES, 65 of the 76 samples (86%) show a negative association, 4 samples (5%) show a positive association, and the remaining 7 cases (9%) are indeterminate. For African BGA-SES associations, 63 of the 77 samples (82%) show negative estimated associations, 10 samples (13%) show a positive association, and the remaining 4 samples (5%) are indeterminate. In all three cases the signed proportions are highly statistically significant against the null hypothesis of no underlying association.

The indirect methodology adopted in this paper for measuring the correlation/association between BGA and SES from epidemiological studies is not coincidental. Research using genotyped DNA is limited by data cost and availability, and there is potential political backlash

against findings linking BGA to any socially desirable trait such as intelligence or SES. The epidemiological studies meta-analyzed in this paper show clear and consistent links between BGA and SES but do so tangentially; they perform the analysis because measuring the linkages between explanatory variables is a standard monitor on regression model reliability. These studies are protected from the usual backlash against politically sensitive findings since their examination of the linkage between BGA and SES is only undertaken as a peripheral check on regression model stability and reliability. Nonetheless, examining a broad swath of these studies, the meta-analytic results are clear and consistent despite not being the empirical focus in any of the individual studies.

Analyzing SES has long been a major research topic in economics, yet the recent successful use of BGA admixture proportions in epidemiology has not been reflected in economic modeling. As an example, the highly regarded research center Opportunity Insights at Harvard University has produced 26 research papers (as of June 2024) using vast quantities of data from a wide range of sources to examine numerous aspects of SES and its dynamic cross-sectional distribution in the US. Not one of the Opportunity Insights papers utilizes BGA admixture proportions data. Admittedly, admixture proportions data is relatively cumbersome and expensive since it requires DNA sampling and genotyping, and it is also politically sensitive since it touches upon group genetic variation and racial SES gaps. Nonetheless, given that the recent epidemiology research literature indirectly shows clear and consistent patterns of linkage between SES and BGA admixture proportions, this potentially powerful new data source deserves careful consideration, or a detailed explanation for its exclusion in the study of SES by economists. There is a serious risk of confounding if the BGA-SES correlation is omitted in economic analysis of SES.

II. Data

A. Study identification, screening, and selection

We created a database of all published epidemiological studies for which associations between continental-level biogeographic ancestry and socioeconomic outcomes were reported, limited to those using sample data exclusively from the Americas. Each of the studies included in the meta-analysis incorporate admixture proportions from at least two of the three biogeographic groups African, Amerindian, or European, a socioeconomic status index or a component of such an index, and some statistical measurement of the association between them.

First, we incorporated data from Kirkegaard et al. (2017), who conducted a systematic review of the literature up to 2016 using searches of the PubMed, BIOSIS, and Google Scholar databases. Second, we conducted new searches in PubMed, BIOSIS, and Google Scholar for dissertations or articles written in English, Spanish, or Portuguese and published between 2017 and 2023. We scanned all BIOSIS and PubMed abstracts and scanned the first 1,500 Google Scholar abstracts (ranked by relevance). The searches we employed were:

1. PubMed: (admixture OR genetic ancestry OR genomic OR biogeographical) AND (socioeconomic OR education OR income OR SES OR poverty) AND (African OR European OR Amerindian)

2. BIOSIS: (admixture OR genetic ancestry OR genomic OR biogeographical) AND (socioeconomic OR education OR income OR SES OR poverty) AND (African OR European OR Amerindian)

3. Google Scholar: (admixture OR genetic ancestry) AND (socioeconomic OR education OR income OR SES OR poverty) AND (African OR European OR Amerindian OR Native American)

Note, since Google Scholar yielded many more hits than PubMed or BIOSIS, we slightly altered the terms to provide a more tailored search.

Two of the authors reviewed the paper abstracts for discussion of genetic ancestry in relation to socioeconomic status and coded them accordingly. Codings were compared and discussed until consensus was reached, prior to reading the full articles. We adopted the following criteria for inclusion in the meta-analysis:

1) Samples were from the Americas

2) Samples had an average admixture % of greater than 3.125% (or 1/32nd) for at least two of the following three BGAs: European, African, and Amerindian.

3) Directions of associations between European, African, or Amerindian genetic ancestry and socio-economic status were reported.

4) The reported information was gathered at the individual rather than the group level (e.g., not state-level admixture and state-level social outcomes).

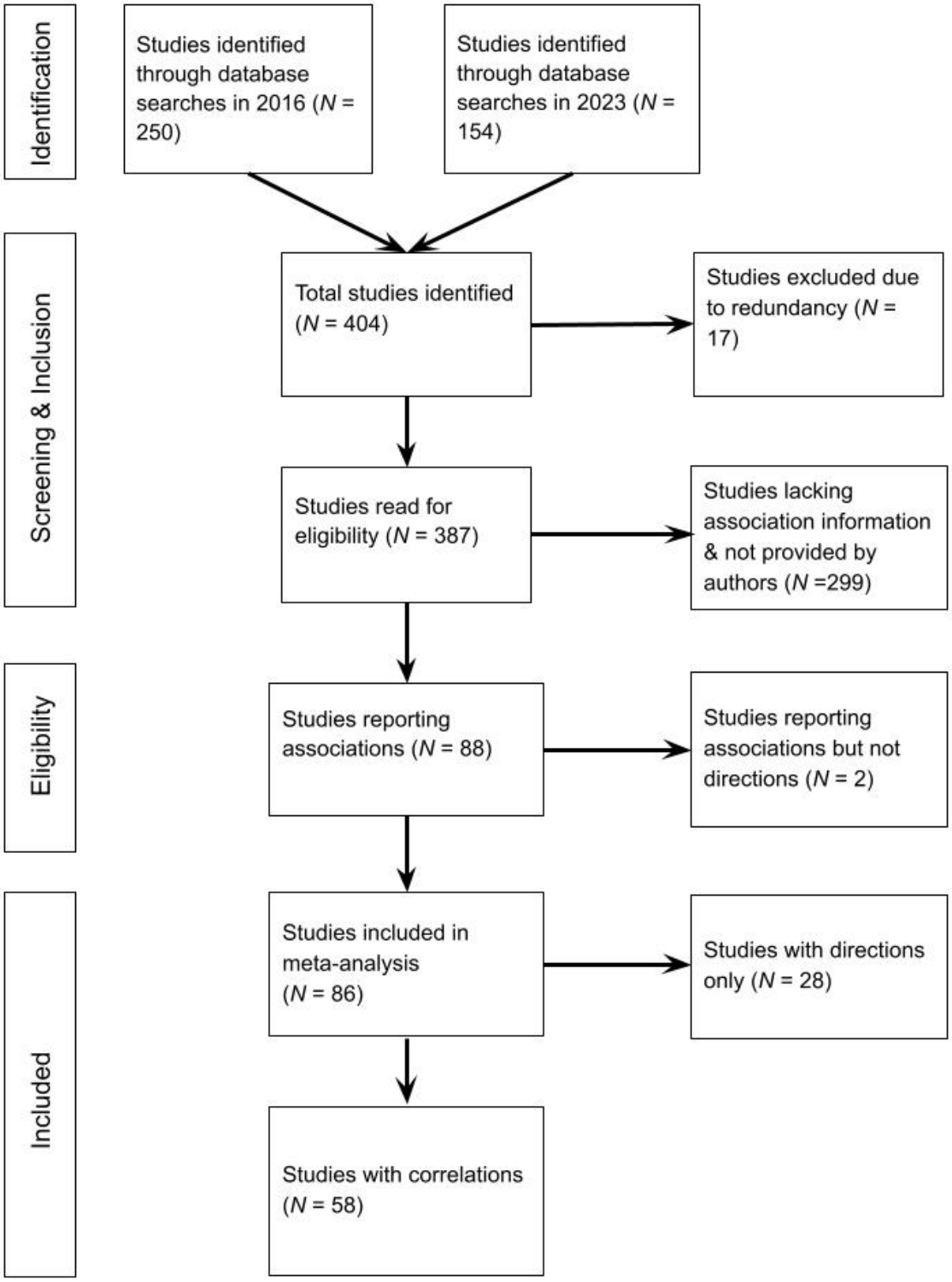
5) Samples did not contain redundant material. In cases of redundant or partially redundant samples, we selected those with the largest sample sizes and most complete information.

6) The study was published after the year 2000.

When relevant data was available, we copied or computed Pearson correlation coefficients where possible. When directions of association were reported but statistics sufficient to compute correlations were not, we emailed the corresponding authors for data. Thirty-one research teams were contacted for the 2023 round of data collection, whereas 26 teams were contacted for the 2016 round. We additionally scanned papers for references reporting associations and then examined these referenced papers. Three out of 31 research teams provided data in the 2023 round, whereas 11 out of 26 authors previously provided data in the 2016 round.

We located a total of 404 studies, including 154 studies from the 2023 search. Out of the 404, seventeen studies were omitted because they had redundant samples. Of the 387 remaining, 299 did not report associations between BGA and socioeconomic indexes. Additionally, there were two studies that identified an association but did not specify the direction of the relationships. The remaining 86 studies provided information on the direction of the association, or such information was given upon request by the authors. Of these, 58 studies included information on the estimated correlation (or related statistics that could be converted into estimated correlation). Figure 1 displays a flowchart of the study selection process. Full details on the studies and reasons for exclusions are reported in Supplementary Materials A.

Figure 1: Flowchart of study inclusion



All studies included the number of observations. In a few instances, owing to missing data, the same sample had different sample sizes for correlation of BGA with different socioeconomic outcomes. In these instances, we recorded the sample sizes specific to the outcome. Additionally, some studies provided data for multiple overlapping samples (e.g., African-American, Hispanic, and White Americans vs. a combined sample). In these cases, we retained specific ethnic groups instead of combined multi-ethnic samples and we retained the ethnic category most consistent across all samples. Furthermore, many types of socioeconomic outcomes were reported; we condensed these into two very broad groups: ones including general SES or composites of multiple indicators such as education, income, and assets, and ones including only individual SES indicators. Prior to conducting the meta-analyses we visually examined the data for outliers. There was a very clear outlier. Specifically, Klimentidis et al. (2009) reported a correlation of r = -.59 between European ancestry and income for a sample of fifteen socially self-identifying Native Americans in the US, while also reporting a correlation of r = .12 between European ancestry and parental education for the same sample. Given the small sample size this estimate is not implausible due to sample variation; as such we retained the data point.

B. Derivation of correlations from equivalent reported statistics

As is commonplace in meta-analytic studies, not all the surveyed studies directly reported the statistic of interest (Pearson correlation coefficient which we denote r) but often reported an equivalent statistic or set of statistics. In this subsection we describe the simple methods we use to infer estimated correlations in this situation.

A total of 157 of the estimates were Pearson correlation coefficients. Spearman correlation coefficients based on comparative rankings were reported for 10 cases, all of which had sample sizes greater than 90. Koricheva et al. (2013) show that when the sample size is greater than 90, Spearman correlation and Pearson correlation are approximately equal under broad circumstances; for this reason, we treated the Spearman estimates as equivalent to r. In 32 cases, authors reported the R-squared from the univariate linear regression of SES on BGA. The square root of the reported R-squared multiplied by the reported sign of the relationship is the estimated correlation. In 13 cases, the authors report the standardized regression coefficient $\hat{\beta}$ from the univariate regression of SES on BGA, which we take as equivalent to the Pearson correlation coefficient. In one case, the authors report the sign of the estimated correlation and its

p-value. We invert the *p*-value by assuming a normal distribution was used for the *p*-value computation. In 57 cases, correlations were not reported but the authors provide k-by-j tables of data frequencies jointly sorted into admixture proportion and SES quantiles. The k-by-j frequencies were converted into *r* using the formula detailed in Fagerland et al. (2017).

C. Directions of association

Reported statistics sufficient to compute the directions of associations between BGA admixture proportions and SES were reported more frequently than those required to compute correlations. An analysis of directions can provide insight into whether results are consistent with a null hypothesis, which in this context posits that a given ancestry component will not be positively or negatively associated with socioeconomic outcomes. Such a method for analyzing the directions of associations has been utilized in previous meta-analyses, exemplified by the work of Van der Meer and Tolsma (2014). For each sample, associations were categorized as either negative, null, or positive, where 'null' indicates either no association or in the case of multiple measured associations using the same sample no majority finding of either positive or negative associations. This categorization did not take into account either the statistical significance or the strength of the association.

III. Empirical Findings

A. Descriptive statistics of the study samples

Most of the papers analyzed were published in recent years, with the median publication year being 2015, spanning from 2002 to 2024. Only a subset of the independent samples reported associations for all three BGAs; moreover, only a subset of samples with directions of associations also had correlations or equivalents. As a result, the number of samples with associations for a particular BGA is less than the total number of samples and the number of samples with correlations is less than that with directions of associations. Table 1 displays the number of samples by BGA and association type.

Table 1. Number of samples, estimates and individual observations by biogeographic ancestry (BGA) and by relationship type

Relationship type	Observation count type	Any BGA	European BGA	Amerindian BGA	African BGA
Correlations or equivalents	Samples	88	55	59	57
	Estimates	270	82	94	94
	Individuals	127463	94768	69157	97493
Directional Associations	Samples	117	68	76	77
	Estimates	372	105	119	148
	Individuals	149764	100668	83287	116065

Note: 'Any BGA' refers to the total sample not disaggregated by BGA. The number of samples and individuals for Any BGA is larger than that for an individual BGA but smaller than the sum of the three BGAs because individual samples could contain data for one to three BGAs. "Estimates" refers to the number of estimates computed; this number is greater than the number of samples since in some instances the same sample had multiple estimates.

For the 88 independent samples that had correlations or equivalents and the 117 independent samples that had directions, there were, respectively, a total of 270 and 372 estimates of the association between BGA and outcomes, since multiple estimates were frequently reported for the same samples. The discrepancy between the number of independent samples and the number of estimates arose because some samples contained data for multiple BGAs and/or multiple SES indicators. Sixty out of the 117 independent samples that provided directional data, and 52 out of the 88 independent samples that provided correlations, originated from the US alone or in combination with Puerto Rico. The remaining samples came from Latin America, including (with the number of associations): Brazil (12), Puerto Rico (8), Chile (8), Mexico (8), Colombia (7), Peru (5), and Uruguay (2), and any other Latin American country (7).

Many of the studies in our database provide BGA-SES correlations estimated exclusively using individuals who have chosen the self-identified race or ethnicity (SIRE) category African American, and other studies exclusively for the SIRE category Hispanic. All the SIRE-restricted estimates came from US-only data. Table 2 shows the breakdown of correlation estimates by the SIRE category restrictions.

Table 2. Number of correlation estimates with self-identified race and ethnicity (SIRE) exclusion criteria in sampling and with non-exclusive sampling

Sample selection criteria	European BGA-SES correlations	Amerindian BGA-SES correlations	African BGA-SES correlations
African-American only	16	4	29
Hispanic-American only	20	36	16
Non-exclusive	46	54	49

Note: The table shows the number of correlations from samples restricted to self-identified African-Americans (first row), self-identified Hispanic-Americans (second row), and all other samples (third row). All the correlations in the first two rows are from US-only samples.

The studies included a diverse array of SES indicator variables, which we grouped into broader categories. These categories mostly included measures of personal education, income, or household assets. Moreover, several studies included data on parental SES. Given that a large majority of children are biological children, their admixture usually reflects the mean admixture of their parents. Thus, associations between child ancestry and parental SES can be seen as reflecting those between parental ancestry and SES. Additionally, a sizable number of associations were based on neighborhood socioeconomic status instead of individual SES.

Table 3: Number of correlation/association estimates for individual SES component measures

	Correlations or equivalents			Directions of association		
Biographic ancestry /SES component	European BGA	Amerindian BGA	African BGA	European BGA	Amerindian BGA	African BGA
Education	33	41	34	40	47	47
Income	13	18	16	19	22	26
General SES w/Education	12	11	15	13	18	18
Neighborhood SES	8	6	10	12	9	24
Assets	8	9	8	8	9	13
Parental education	4	1	4	7	2	7
General SES, no Education	1	4	0	2	5	1
Occupation	0	0	0	1	3	3
Occupation & income	2	2	2	2	2	3
Health insurance	1	1	2	1	1	3
General SES w/Education	0	1	1	0	1	1
Parental income	0	0	2	0	0	2
Other SES	0	0	0	0	0	1
Total	82	94	94	105	119	148

Note: For each of the thirteen SES measures used across the collection of the studies, the table shows the number of correlation estimates and direction of association estimates that use that particular SES measure.

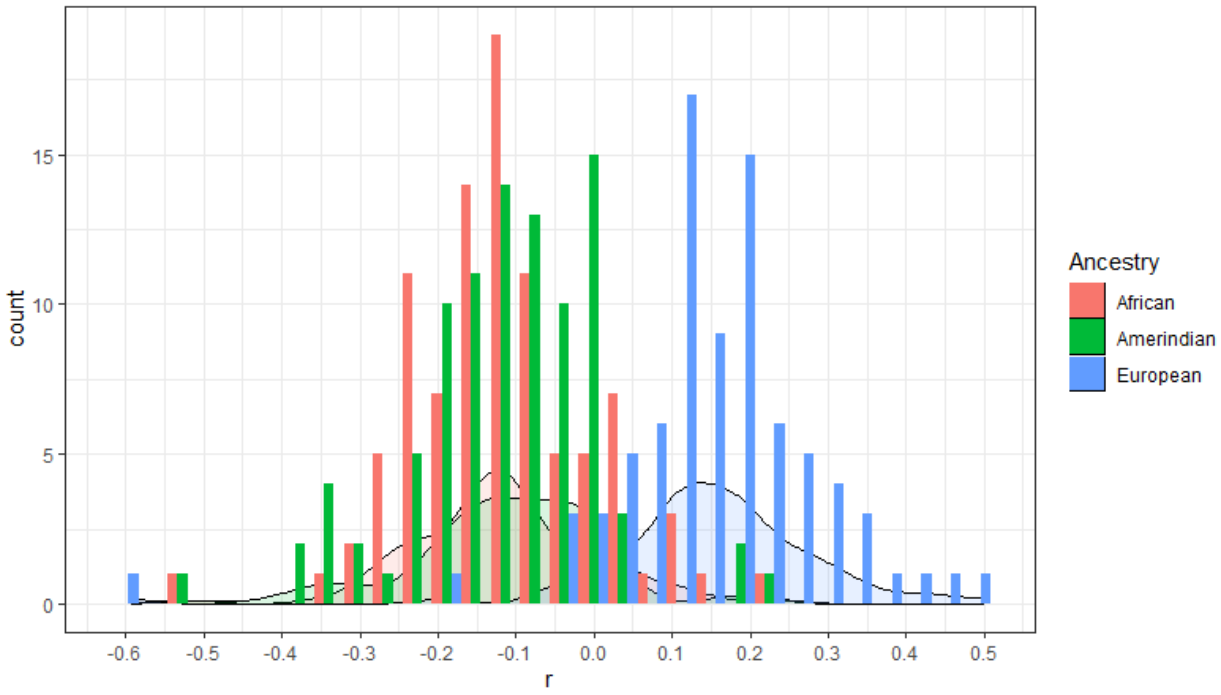
The estimated correlations were mostly in the range (-.25, .30) with a notable tilt by biogeographic ancestry, this tilt will be explored more fully in the next subsection. Table 4 provides some descriptive statistics and Figure 2 shows a histogram color-coded by biogeographic ancestry.

Table 4. Descriptive statistics for estimated correlations (not aggregated within samples)

Biogeographic Ancestry	Number of Estimates	Mean	Median	Max	Min	Standard Deviation	10^{th} centile	90^{th} centile
European BGA	82	.15	.15	.50	-.59	.14	.03	.30
Amerindian BGA	94	-.11	-.10	.23	-.52	.12	-.25	.00
African BGA	94	-.12	-.12	.22	-.52	.11	-.24	.03

Note: The table shows descriptive statistics for the full set of correlations. The statistics are not adjusted for the presence of non-independent estimation error across correlations from the same sample.

Figure 2. Histogram for the estimated correlations (prior to aggregation within samples)



B. Meta-analysis of correlations and directional associations between BGA and SES

We run the meta-analyses separately for each of the three ancestries. For a given BGA, let $r_i,\ i = 1, \ldots, M$ denote our database of M correlation estimates linking the BGA admixture proportion and SES measure. We transform the estimated correlations into Fisher's *Z*-statistics: $Z_i = \frac{1}{2}\ln\left(\frac{1+r_i}{1-r_i}\right)$ to improve the small sample properties of our estimator. We use a random effects model, treating the true Z_i as random across studies and across multiple estimates within studies for those studies which have multiple estimates. We use the correlated and hierarchical effects (CHE) model (Pustejovsky and Tipton, 2022) to account for the fact that estimation errors are correlated within studies with multiple estimates. Letting $\bar{Z}$ denote the expected value of Z_i for all *i,* the CHE model assumes:

$$Z_i = \bar{Z} + \varphi_i + \omega_j + \varepsilon_i$$

where φ_i is the estimate-specific random effect, ω_j is the study-specific random effect associated with estimate *i,* and ε_i is the sample estimation error for this correlation estimate. The index $j = 1, \ldots, K$ runs over the set of samples, some of which have more than one estimate (by estimating the correlation between BGA and several SES measures on the same sample). The CHE model

makes the restrictive assumption that the correlation (ρ) between multiple estimation errors within a common sample has a known constant value. We set this correlation to ρ = .4 (note that the correlation is only used for determining point estimates, not for finding standard errors of these estimates; see below). The CHE model is estimated by restricted maximum likelihood (REML) as recommended by Pustejovsky (2021) and Harrer et al. (2021). See the Technical Appendix for detailed discussion of the model. As a robustness check we also estimated the model using ρ = .8 and found similar results; these supplementary results are shown in Supplementary Materials B.

In order to improve the generality of the restrictive correlation structure of CHE, we enhance the CHE model with the sandwich-estimator-based Robust Variance Estimation (RVE) developed by Pustejovsky and Tipton (2022). This CHE-RVE model uses the simple correlation structure of CHE to solve the likelihood maximization problem, but then computes consistent standard errors of the estimated parameters under much more general conditions on the dependency structure of the within-study estimation errors. See the Technical Appendix or Pustejovsky and Tipton (2022) for a discussion of the CHE-RVE-REML estimation methodology.

The estimation results for each of the three BGAs are shown in Table 5, along with 95% confidence intervals for $\bar{Z}$. In the table we also invert $\bar{Z}$ to re-state it in units of correlation, that is $(\exp\left(\bar{Z} - \frac{1}{2}\right) - 1)/(\exp\left(\bar{Z} - \frac{1}{2}\right) + 1)$. The results in this table are strongly supportive of a nonzero correlation between BGA and SES for each of the three BGAs. Forest plots and funnel graphs of the data are provided in Supplementary Materials B.

Table 5. Estimated average correlations between socioeconomic status measures and biogeographic ancestries

Biogeographic ancestry	Estimated $\bar{Z}$	95% confidence interval for $\bar{Z}$	*P*-value for $\bar{Z} = 0$	Implied correlation from $\bar{Z}$	95% confidence interval for implied correlation
European BGA	.161	[.129, .192]	< .001	.160	[.128, .190]
Amerindian BGA	-.107	[-.145, -.069]	< .001	-.107	[-.144, -.069]
African BGA	-.129	[-.161, -.097]	< .001	-.128	[-.160, -.097]

Note: For each biogeographic ancestry group (European, Amerindian, and African) the correlated hierarchical errors model with robust variance estimation is used to estimate the mean Fisher's *Z* (mean transformed correlation) across all correlation estimates. The table also shows the mean Fisher's *Z* and its confidence interval reverse-transformed into correlations.

The average correlation estimates in Table 5 are not large, but Kirkegaard (2022) demonstrated that very small correlations between BGA and SES can still reflect substantial ancestry effects. For instance, among Chileans, the correlation between European BGA and SES was $r = .13$ ($N = 1805$), but the unstandardized regression slope from 0% to 100% European ancestry was b = 0.88. Fuerst et al. (2024) also observed this phenomenon. This occurs because restricted ancestry ranges lead to attenuated correlations and because bivariate correlations treat BGA and SES associations without a reference BGA category. Thus, our meta-analytic correlations between BGA and SES are consistent with substantial ancestry effects.

As a non-parametric alternative to the meta-analysis of correlations, and to increase the data coverage to include samples where it was not possible to recover a correlation coefficient, we also examine directionally-assigned associations between each BGA and SES. As noted above, the associations found in each sample have been categorized as negative, positive or null. The null category covers all cases where either an individual sample had a measured association of zero or where an individual sample had multiple measures of association and there were an equal number of positive and negative directions (such as having two positive and two negative associations). To convert this trinomial count variable into a binomial test, we use the cautious approach of assigning the null cases to whichever category (positive or negative) has a lower sample proportion and then testing whether the other category has a sample binomial proportion significantly greater than 0.5, thereby biasing the results against rejecting the null hypothesis of no association. This cautious approach lowers the power to reject the null hypothesis but in practice has no impact since we are easily able to reject the null for all three BGA categories.

The results of the association direction analyses are presented in Table 6. For European ancestry, approximately 85% of associations were positive, contrasting with only 3% negative. Similarly, Amerindian ancestry showed 5% positive associations and 86% negative associations, while African ancestry displayed 13% positive associations and 82% negative associations. These highly significant binomial tests indicate deviations from the null hypothesis for all three ancestries.

Table 6. Meta-analytic results for directions of associations between biogeographic ancestries and socioeconomic outcomes

Biogeographic ancestry	Number of individuals in sample	Number of directional associations	% positive	% negative	% null	*P*-value from binomial test of no association
European BGA	100668	68	85.3%	2.9%	18.8%	2.4×10^{-9}
Amerindian BGA	83287	76	5.3%	85.5%	9.2%	1.8×10^{-08}
African BGA	116065	77	13.0%	81.8%	5.2%	1.4×10^{-0}

Note: For each biogeographic ancestry group (European, Amerindian, and African) each sample is assigned an association direction of positive, negative or null between that ancestry component and socioeconomic status measure(s). The *p*-value tests whether the percent positive or negative (whichever is larger) is significantly greater than 50% using a binomial test.

C. Moderator analysis with a mixed-effects model

In this subsection we re-estimate the CHE-RVE model of BGA-SES correlations allowing a conditional expected value $\bar{Z}$ to depend upon a moderator. This is a mixed effects model; see Harrer et al. (2021) or the Technical Appendix for details. First, we use a moderator defining two regional categories: estimates from US data versus those using data from outside the US. The rationale for this moderator is that populations in the US are often relatively recent migrants in the case of Hispanics, or, in the case of White and African-Americans, they do not have as extensive and pervasive a history of exogamy as is found among many Latin American populations. These factors may affect the correlations between BGA and socioeconomic outcomes. The countries for each sample were recorded. For a couple of samples, individuals came from multiple Latin American countries. For one sample, they came from both the US and the territory of Puerto Rico. We coded "region" as "US" if the sample was either from the US or

from both the US and Puerto Rico. The samples were coded as "not US" in all other cases. Since we have one sample from Trinidad & Tobago and one from Dominica, "not US" is not identical to "Latin American". The results for this analysis are shown in Table 7. There is little discernable difference between the correlation estimates in the US versus non-US subsamples; the same highly significant pattern of positive correlation for European BGA-SES and negative correlation for Amerindian and African BGA-SES remains unchanged in all cases.

Table 7: Subgroup analysis based on region (US versus non-US samples)

Moderator	Estimated $\bar{Z}$	95% Confidence Interval	*P*-value for $\bar{Z} = 0$	*P*-value for subgroup equality
European BGA				
US	.155	[.109, .200]	<.001	.647
Non-US	.169	[.123, .214]	<.001	
Amerindian BGA				
US	-.086	[-.146, -.025]	<.001	.278
Non-US	-.126	[-.173, -.079]	<.001	
African BGA				
US	-.134	[-.174, -.093]	<.001	.724
Non-US	-.122	[-.181, -.062]	<.001	

Note: For each of the three BGA groups (European, Amerindian, and African) the table shows the mean Fisher's Z for US and for non-US samples estimated using the correlated hierarchical errors model with robust variance estimation. The final column gives a *p*-value for equality of the two conditional means.

Next, we consider whether the SES-BGA correlations might be mediated by Self-Identified Race or Ethnicity (SIRE). This is because a large body of literature argues that social inequalities are primarily related to socially-defined race and not ancestry (e.g., Adkins-Jackson et al., 2022). We use a trinomial moderator which isolates the three subgroups of estimates shown in Table 2: those from samples restricted to African-Americans, those from samples restricted to Hispanic Americans, and those from all other samples. The results are shown in Table 8. For both African and Hispanic Americans, the associations between BGA and SES are not significantly different from the association in the Other group. Moreover, the positive correlation between SES and European BGA and the negative correlation between SES and African BGA remains significant for all the subgroups. For African-Americans, there is no significant correlation between SES and Amerindian BGA. This is understandable because there is little variance in Amerindian admixture among African Americans. As a result, few authors report associations with Amerindian BGA for this group and so data was available for only 4

samples. Further, range restriction in Amerindian admixture in this group attenuates correlations. Overall, the results indicate that SES-BGA correlations can be found within SIRE groups.

Table 8: Subgroup analysis based on Self-identified Race or Ethnicity (SIRE)

Subgroup	Estimated $\bar{Z}$	95% Confidence Interval	*P*-value for $\bar{Z}$	*P*-value for subgroup equality
European BGA				
African-American (US)	.132	[.071, .192]	.002	.199
Hispanic-American (US)	.167	[.071, .264]	.006	.848
Non-exclusive	.176	[.130, .222]	< .001	
Amerindian BGA				
African-American (US)	-.052	[-.357, .254]	.277	.253
Hispanic-American (US)	-.088	[-.161, -.016]	.021	.393
Non-exclusive	-.123	[-.167, -.079]	< .001	
African BGA				
African-American (US)	-.124	[-.163, -.084]	<.001	.505
Hispanic-American (US)	-.099	[-.197, -.001]	.049	.336
Non-exclusive	-.148	[-.212, -.083]	< .001	

Note: For each of the three BGA groups (European, Amerindian, and African) the table shows the mean Fisher's *Z* for samples restricted to African-Americans, to Hispanic-Americans, and to all other (that is, non-exclusive) samples. The model is estimated using the correlated hierarchical errors model with robust variance estimation. The final column gives a *p*-value for equality of the restricted-samples conditional mean to the non-exclusive-samples conditional mean.

Next we test if the relationship between BGA and socioeconomic outcomes might vary depending on how well the SES indicator measures relevant facets of SES. This is done because a large study reported that associations with genetic ancestry were more pronounced on comprehensive measures of SES (Fuerst et al., 2024). Therefore, we consider SES-indicator type (SES index versus SES component) as a moderator variable. We recorded the SES type (e.g., income, education, health insurance, neighborhood SES, general SES w/Education). An SES type was coded as general SES index if it was a composite of multiple SES indicators (e.g.,

Hollingshead index). We recoded these categories as “general SES index” if the outcome involved general SES (i.e., "General SES, no Education", "General SES w/Education", "Parental General SES w/Education", or "Parental General SES, no Education"). The results are shown in Table 9. Based on the point estimates, in two of the three cases the SES composites produce marginally stronger BGA-SES correlations than SES indicators, but the difference is never statistically significant. In the case of African BGA, the point estimates are virtually identical for composite versus indicator SES measures. Thus, we do not find significant evidence that composite measures of SES are more strongly related to BGA relative to single indicators.

Table 9: Subgroup analysis based on socioeconomic status indicator type

Moderator	Estimated $\bar{Z}$	95% Confidence Interval	*P*-value for $\bar{Z} = 0$	*P*-value for subgroup equality
European BGA				
SES indicator	.142	[.104, .181]	< .001	.064
SES composite	.230	[.141, .319]	<. 001	
Amerindian BGA				
SES indicator	-.090	[-.144, -.036]	.002	.078
SES composite	-.181	[-.249, -.112]	< .001	
African BGA				
SES indicator	-.129	[-.169, -.090]	< .001	.976
SES composite	-.128	[-.179, -.078]	< .001	

Note: For each of the three BGA groups (European, Amerindian, and African) the table shows the mean Fisher’s *Z* for single-component SES measures and for composite SES measures; the model is estimated using the correlated hierarchical errors model with robust variance estimation. The final column gives a *p*-value for equality of the two conditional means.

Finally, we examine whether there might be differences due to the geographic distribution of the sample (national or international vs. local); 55% of the samples were local, meaning that individuals came from either the same city or the same first-order administrative division within a country (e.g., state or province). Individuals in the remaining 45% of samples came either from multiple administrative divisions within a country or multiple countries. Data was categorized as local if it fell within a first-order administrative unit (FOAD) (e.g., US state or territory) within a medium size to large country or it fell within a local region within a very small country (even if that region included multiple FOADs). There were only three small-country cases: Dominica, Central Valley of Costa Rica and Northern Trinidad of Trinidad &

Tobago. It made little sense to treat Puerto Rico as local, but then Dominica as national/international, so we made these three exceptions and treated these three small-country cases as local. The results are shown in Table 10. There is no significant evidence of differences between these subsamples.

Table 10: Subgroup analysis based on geographical distribution of the sample (local versus multiple-region samples)

Moderator	Estimated $\bar{Z}$	95% Confidence Interval	P-value for $\bar{Z} = 0$	P-value for subgroup equality
European BGA				
Local	.197	[.149, .246]	<.001	.265
Multiple region	.161	[.117, .205]	<.001	
Amerindian BGA				
Local	-.134	[-.176, -.091]	<.001	.445
Multiple region	-.110	[-.156, -.065]	<.001	
African BGA				
Local	-.129	[-.176, -.082]	<.001	.836
Multiple region	-.122	[-.173, -.072]	<.001	

Note: For each of the three BGA groups (European, Amerindian, and African) the table shows the mean Fisher's Z for local samples and for multiple-region samples; the model is estimated using the correlated hierarchical errors model with robust variance estimation. The final column gives a p-value for equality of the two conditional means.

V. Summary

This paper performs a meta-analysis of estimated correlation coefficients linking African, Amerindian or European biogeographic ancestry to socioeconomic status. The estimated correlation coefficients collected in the meta-analysis originate from epidemiological studies. Many epidemiological studies include both biogeographic ancestry admixture proportions and socioeconomic status measures as explanatory variables in studying health traits, and it is standard procedure to estimate the correlation coefficients between these key explanatory variables as a monitor on statistical model under-identification or misspecification. None of the surveyed studies are directly concerned with estimating this correlation but rather do so for statistical monitoring of model reliability while focused on some other topic. By performing the same statistical quality control across diverse countries and samples, these epidemiological studies have inadvertently created a powerful body of evidence showing correlation between biogeographic ancestry and socioeconomic status.

European BGA admixture proportion shows a significantly positive correlation with SES, whereas both Amerindian and African BGA admixture proportions are significantly negatively correlated with SES. The same statistically significant patterns emerge in the directionality of the BGA-SES associations: European BGA admixture proportion has a positive association with SES whereas both Amerindian and African BGA admixture proportions have a negative association; all three of these signed patterns of association are highly statistically significant.

The observed correlations are not explained by self-identified race or ethnicity (SIRE) since correlation estimates on samples restricted to African-American SIRE and to Hispanic-American SIRE show essentially the same patterns as unrestricted samples. Also, it is not explained by particularly American historical or institutional features since the pattern of correlations is very similar in US samples and non-US samples.

The correlations between BGA admixture proportions and SES are a tangential concern in epidemiology but have core relevance to the economic study of socioeconomic status. Since this meta-analysis clearly shows that BGA admixture proportions have significant univariate correlation with SES, it is incumbent to consider BGA admixture proportions as a candidate variable in economic analysis of SES. Obtaining this variable can be cumbersome and expensive since it requires DNA sampling and genotyping, but its complete absence from extant results makes the existing economic analysis of SES unreliable. There is a potential omitted-variable bias of notable concern.

References

Adkins-Jackson, P. B., Chantarat, T., Bailey, Z. D., & Ponce, N. A. (2022). Measuring structural racism: a guide for epidemiologists and other health researchers. American Journal of Epidemiology, *191*(4), 539-547.

Anonymous (2024). Opportunity Insights, List of all papers. Harvard University, https://opportunityinsights.org/paper/

Becker, N. S., Verdu, P., Froment, A., Le Bomin, S., Pagezy, H., Bahuchet, S., and Heyer, E. (2011). Indirect evidence for the genetic determination of short stature in African Pygmies. *American journal of physical anthropology,* 145(3), 390-401.

Bidulescu, A., Choudhry, S., Musani, S. K., Buxbaum, S. G., Liu, J., Rotimi, C. N., … and Gibbons, G. H. (2014). Associations of adiponectin with individual European ancestry in African Americans: the Jackson Heart Study. *Frontiers in genetics*, 5, 22.

Connor, G., J. Fuerst, and M. Hu (2024). Supplementary Materials A for "Correlations between socioeconomic status and biogeographic ancestries": Data file and list of included papers, Open Science Framework, https://osf.io/39hbd

Connor, G., J. Fuerst, and M. Hu (2024). Supplementary Materials B for "Correlations between socioeconomic status and biogeographic ancestries": Additional tables and figures, Open Science Framework, https://osf.io/erdyj

Fagerland, M., S. Lydersen, and P. Laake (2017) Statistical analysis of contingency tables, Chapman and Hall Biostatistics Series, Routledge Publishers, New York.

Fernández, J.R. and Shiver, M.D. (2004). Using genetic admixture to study the biology of obesity traits and to map genes in admixed populations. *Nutrition Reviews 62*(S2): S69-S74.

Flores, C., Ma, S. F., Pino-Yanes, M., Wade, M. S., Pérez-Méndez, L., Kittles, R. A., … and Garcia, J. G. (2012). African ancestry is associated with asthma risk in African Americans. *PloS one*, 7(1), e26807.

Fuerst, J. G., Shibaev, V., and Kirkegaard, E. O. (2024). A Genetic Hypothesis for American Race/Ethnic Differences in Mean g: A Reply to Warne (2021) with Fifteen New Empirical Tests Using the ABCD Dataset. In: G. Connor and J. Fuerst (Eds). *Correlations between Genetic Variation and Test Score Gaps*. Cambridge Scholars Publishing.

Halder, I., Matthews, K. A., Buysse, D. J., Strollo, P. J., Causer, V., Reis, S. E., and Hall, M. H. (2015). African genetic ancestry is associated with sleep depth in older African Americans. *Sleep*, *38*(8), 1185-1193.

Harrer, M., Cuijpers, P., Furukawa, T. and Ebert, D. D. (2019). *dmetar: Companion R Package For The Guide 'Doing Meta-Analysis in R'*. R package version 0.1.0.

Harrer, M., Cuijpers, P., Furukawa, T.A., and Ebert, D.D. (2021). *Doing Meta-Analysis with R: A Hands-On Guide*. Boca Raton, FL and London: Chapman and Hall/CRC Press.

Hedges, L.V., Tipton, E. and Johnson, M.C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. Research synthesis methods, Vol. 1, 39-65.

Kirkegaard, E. O. (2022). Admixture and social status in Chile. *Mankind Quarterly, 62*(4).

Kirkegaard, E. O. W., Wang, M., and Fuerst, J. (2017). Biogeographic Ancestry and Socioeconomic Outcomes in the Americas: A Meta-Analysis. *Mankind Quarterly, 57*(3), 398–427.

Klimentidis, Y. C., Miller, G. F., and Shriver, M. D. (2009). The relationship between European genetic admixture and body composition among Hispanics and Native Americans. *American Journal of Human Biology: The Official Journal of the Human Biology Association*, *21*(3), 377-382.

Koricheva, J., J. Gurevitch, and K. Mengersen (2013). *Handbook of meta-analysis in ecology and evolution*, Princeton University Press, New Jersey, USA.

Langan, D., J.P.T. Higgins, …, M. Simmons (2018) A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses. Research Synthesis Methods.

Lenhard, W. and Lenhard, A. (2022). Computation of effect sizes. Psychometrica. Retrieved from: https://www.psychometrica.de/effect_size.html.

Mehanna 2021

Olkin, Ingram, and Jeremy D Finn. 1995. "Correlations Redux." Psychological Bulletin 118 (1): 155.

Parcha, V., Heindl, B., Kalra, R., Bress, A., Rao, S., Pandey, A., ... and Arora, P. (2022). Genetic European ancestry and incident diabetes in Black individuals: Insights from the SPRINT trial. *Circulation: Genomic and Precision Medicine*, *15*(1), e003468.

Pustejovsky, J. E., and Tipton, E. (2022). Meta-analysis with robust variance estimation: Expanding the range of working models. *Prevention Science, 23*(3), 425-438.

James E. Pustejovsky (2021). Variance component estimates in meta-analysis with mis-specified samping correlation Published November 28th 2021, White Rose Publishing Online, https://jepusto.com/posts/variance-components-with-misspecified-correlation/

Rhead, B., Hein, D., Pouliot, Y., Guinney, J., De La Vega, F., and Sanford, N. N. (2022). Genetic ancestry differences in tumor mutation in early and average-onset colorectal cancer.

Shibaev, V., and Fuerst, J. (2024). A genetically informed test of the cognitive-colorism hypothesis. In: G. Connor and J. Fuerst (Eds). *Correlations between Genetic Variation and Test Score Gaps*. Cambridge Scholars Publishing.

Tipton, E. and J.E. Pustejovsky (2015). Small-sample adjustments for tests of moderators and model fit using robust variance estimation in meta-regression. Journal of educational and behavioral studies, Vol. 40: 604-634.

Van den Noortgate, W., López-López, J. A., Marín-Martínez, F., and Sánchez-Meca, J. (2013). Three-level meta-analysis of dependent effect sizes. *Behavior research methods, 45*, 576-594.

Van der Meer, T. and Tolsma, J. (2014). Ethnic diversity and its effects on social cohesion. *Annual Review of Sociology* 40: 459–478.

Wolfgang Viechtbauer (2005). Bias and efficiency of meta-analytic variance estimators in the random-effects model, Journal of educational and behavioral statistics Fall 2005 Vol. 30, no. 3 pp. 261-293.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1-48. https://doi.org/10.18637/jss.v036.i03\

Wilson, D. B. (2023). Practical meta-analysis effect size calculator (Version 2023.11.27). https://www.campbellcollaboration.org/escalc/html/EffectSizeCalculator-Home.php

Zuo, L., Luo, X., Listman, J. B., Kranzler, H. R., Wang, S., Anton, R. F., ... and Gelernter, J. (2009). Population admixture modulates risk for alcohol dependence. *Human genetics*, 125(5), 605-613.

Technical Appendix: Review of CHE-RVE-REML

The paper uses the **R** programming language, **metaphor** library, routine **rma.mv** to estimate the CHE-RVE-REML model without moderators (Table 5) and with four moderators each used singly (Tables 7-10). This short appendix reviews this methodology and gives some details of our implementation.

In our application the model starts with the meta-analytic database of estimated correlations, $r_i, i = 1, \dots, M$ with numbers of observations in the original source studies of $N_i, i = 1, \dots, M$. As recommended by Harrer et al. (2021) our meta-analysis of estimated correlations uses as input the transformed values of the estimated correlations into Fisher's Z-statistics:

$$Z_i = \frac{1}{2}\ln\left(\frac{1+r_i}{1-r_i}\right)$$

which are treated as multivariate normal with known estimation variances.

The restricted maximum likelihood (REML) approach to meta-analytic model estimation is well-established as a favored technique. For purposes of estimating the variance components it involves maximizing the likelihood on a reduced sample after restriction of the original set of observations to contrasts: each observation is replaced with its original value minus a fixed linear combination of the observations. This shrinks the number of linearly independent observations in the empirical likelihood but improves the small-sample properties of the resulting variance estimates. The REML estimate of the mean(s) uses the raw observations not the contrasts. In a simulation comparison of estimators for meta-analysis with random effects models,Viechtbauer (2005) concludes "the restricted maximum likelihood estimator strikes a good balance between unbiasedness and efficiency and, therefore, could be generally recommended." Langan et al. (2018) use simulation methods to compare various estimators and recommend REML for estimating random effects models; Pustejovsky (2021) specifically recommends REML for estimation of the CHE-RVE model. We use REML as implemented in the **rma.mv** routine via the option **Method = REML**.

The CHE-RVE model originates with Putejovsky and Tipton (2021); see Harrer et al. (2021, ch. 10) for **R** implementation guidelines. The CHE-RVE model is motivated by meta-analytic applications like ours which need to allow for non-independent sampling errors in the database

of study estimates. Many of the studies in our database include multiple estimates of BGA-SES correlations for a given BGA. For the particular BGA being analyzed let $j = 1, \dots, S$ denote the set of independent samples $S \leq M$ with the number of estimates in each sample $K_j, j = 1, \dots, S$ where $K_j \geq 1$. In our meta-analytic database some of the estimates come from samples with only one correlation estimate for that BGA whereas others have multiple estimates using different SES measures; see Table 1 in the paper for *M* and *S* for each BGA.

The CHE-RVE model begins with a three-level model of the randomness in Z_i:

$$Z_i = \bar{Z} + \varphi_i + \sum_{j=1}^{S} D_{ij}\omega_j + \varepsilon_i \;\; i = 1, \dots, M$$

where $\bar{Z}$ is the mean to be estimated, ε_i is the (level-1) estimation error of estimate *i*, φ_i is the (level-2) individual-estimate level random effect, and ω_j is the (level-3) sample-level random effect. The dummy variable D_{ij} equals one if estimate *i* comes from sample *j* and zero otherwise. The random effects φ_i, ω_j are mutually independent for all *i,j* with $\varphi_i \sim N(o, \sigma_\varphi^2)$, $\omega_j \sim N(o, \sigma_\omega^2)$ and independent of ε_i for all *i,j*. The known variance of $\varepsilon_i, \sigma_{\varepsilon i}^2$ is assumed to equal $\frac{1}{N_i - 3}$ for each *i*; Olkin and Finn (1995) show that this well approximates the true variance for reasonably large N_i.

The estimation errors are assumed independent if they come from different samples, but if $\varepsilon_i, \varepsilon_{i*}$ come from the same sample *j* then they are assumed to have fixed (known) correlation ρ. This simple structure imposed on the covariance matrix of estimation errors allows Putejovsky and Tipton (2021) to solve the REML problem elegantly. We impute the full block-diagonal covariance matrix of $\varepsilon_i, i = 1, \dots, M$ from ρ and $\sigma_{\varepsilon i}^2, i = 1, \dots, M$ using the **R** routine **impute_covariance_matrix**.

Putejovsky and Tipton (2021) note that this simple correlation restriction on the covariances of the within-sample estimation errors is necessary for their solution to the likelihood maximization problem but it is not necessary to impose this assumption when computing the parameter estimation variances. They propose a Huber-White sandwich estimator for the parameter estimation variances. In the "middle term" of the Huber-White sandwich estimator the covariance submatrix within the same-sample diagonal blocks is proxied by the outer product of the demeaned observations; see Tipton and Pustejovsky (2015, p. 608). Outside these block

diagonals the covariance matrix is set to zero. We implement this in R using **vcov = "CR2"** in **metaphor** with the **clubsandwich** library.

Tables 7-10 in the paper enhance the analysis done in Table 5 by including a moderator, giving a mixed effects model. The variance parameters are assumed constant across the subgroups and only the means differ. Let D_i^m denote a dummy variable which is one if estimate *i* comes from the moderator subgroup and zero otherwise. The model becomes:

$$Z_i = \overline{Z_0}(1 - D_i^m) + \overline{Z_1}D_i^m + \varphi_i + \sum_{j=1}^{S} D_{ij}\omega_j + \varepsilon_i \;\; i = 1, \ldots, M$$

Except for the presence of two mean parameters $\overline{Z_0}$, $\overline{Z_1}$, the modeling assumptions and estimation methodology are unchanged. In Table 8 we also consider the case with three subgroups, replacing $\overline{Z_0}(1 - D_i^m) + \overline{Z_1}D_i^m$ with $\overline{Z_0}D_i^{m0} + \overline{Z_1}D_i^{m1} + \overline{Z_2}D_i^{m2}$ with the three dummy variables defined in the obvious way.