Linear and partially linear models of behavioural trait variation using admixture regression

Gregory Connor*
Maynooth University

Bryan J. Pesta
Cleveland State University

June 13, 2021

Abstract

Admixture regression methodology exploits the natural experiment of random mating between individuals with different ancestral backgrounds to infer the environmental and genetic components to trait variation across culturally-defined racial and ethnic groups. This paper provides a new statistical framework for admixture regression based on the linear polygenic index model widely used in behavioural genetics. Using this framework we develop a new test of the differential impact of multi-racial identities on trait variation, an orthogonalization procedure for added explanatory variables, and a partially linear semiparametric functional form. The methodology is illustrated with data from the Adolescent Brain Cognitive Development Study.

Keywords: admixture regression, linear mixed effects model, partially linear semiparametric regression, orthogonalized regressors.

*Corresponding author. Correspondence should be addressed to Gregory Connor, School of Business, Maynooth University, Maynooth, Co. Kildare, Ireland (email: gregory.connor@mu.ie).

1 Introduction

Racial/ethnic group identities such as Black, White, Hispanic, Native American, East Asian and South Asian show empirically strong linkages to medical and behavioural traits such as obesity (Wang et al. 2007), type 2 diabetes (Cheng et al. 2013), hypertension (Lackland 2014), asthma (Choudry et al. 2006), neuropsychological performance (Llibre-Guerra et al. 2018), smoking behaviours (Choquet et al. 2021), and sleep disorders (Halder et al 2015). An important research question is to what degree any such observed trait variation arises from differences in the typical diets, cultural practices and other environmental particularities of the racial/ethnic groups, or from similarity in genetic pools within each group traceable to shared geographic ancestry. Many diverse national populations descend demographically from isolated continental groups within a few hundred years. Modern genetic technology can measure with high accuracy the proportion of an individual's ancestry associated with these continental groups. Also, in many culturally diverse nations, most individuals can reliably self-identify as members of one or more racial or ethnic groups. Admixture regression leverages these two data sources, self-identified race or ethnicity (SIRE) and genetically-measured admixture proportions, to decompose trait variation correspondingly. Admixture regression has been widely applied to medical and behavioural traits including asthma (Salari et al. 2007), body mass index (Klimentidis et al. 2009), type 2 diabetes (Cheng et al. 2013), blood pressure (Klimentidis et al. 2012), neuropsychological performance (Lasker et al. 2019), and sleep depth (Halder et al. 2015). It has particular value in the case of complex behavioural traits where reliably identifying genetic loci associated with trait variation is beyond the current reach of science. Admixture mapping is a more technically challenging methodology, often used in conjuction with admixture regression, which uses ancestral population trait differences to attempt to identify genetic loci associated with a trait. This paper focusses exclusively on admixture regression.

This paper first develops a new statistical framework for admixture regression of behavioural traits by linking it to the linear polygenic index model from behavioural genetics; this framework clarifies the key assumptions that are implicit in this simple and powerful statistical technique. The paper then extends the admixture regression methodology in several ways. We provide a new test statistic for identifying whether a given multi-racial identity differs in its trait impact from the average impact of its component single-SIRE

categories. We examine the role of additional explanatory variables in the admixture regression and their interpretation with and without orthogonalization with respect to the core explanatory variables. We generalize the linear admixture regression specification to a partially linear semiparametric form.

We illustrate our methodology using neuropsychological performance data from the Adolescent Brain Cognitive Development database. Neuropsychological performance is one of the most complex traits to which admixture regression analysis has been applied. Using our new test statistic, we find that some multi-racial categories have identifiably distinct impact on trait variation relative to their component categories. We find that orthogonalization of additional variables can substantially change the interpretation of the core coefficients in the admixture regression. Our results hint that a partially linear semiparametric specification potentially adds empirical value.

2 A statistical framework for admixture regression tests of trait variation

2.1 Variable definitions

We assume that the database consists of n individuals indexed by $i = 1, n$ who have each self-identified their racial or ethnic group membership(s), recorded a score on a behavioural trait, s_i, and provided a personal DNA sample. The k racial or ethnic group self-identification choices are captured by a matrix of zero-one dummy variables $SIRE_{ij}$, $i = 1, n$; $j = 1, k$. We assume that every individual has self-identified as belonging to at least one and possibly more of the k groups.

We assume that a set of m geographic ancestries covered in the study have been chosen, such as African, European, Amerindian, South Asian, and East Asian, indexed by $h = 1, m$. The genotyped DNA samples are carefully decomposed into admixture proportions of geographic ancestry, as discussed in Section 4 below. For each individual the ancestry proportions across the chosen geographic ancestries sum to one. This gives a matrix of ancestry proportions A_{ih}, $i = 1, n$; $h = 1, m$ with $0 \leq A_{ih} \leq 1$ for all i, h and $\sum_{h=1}^{m} A_{ih} = 1$ for each i.

In most applications of admixture regression, individuals' racial or ethnic

group identities will have statistical relationships with individuals' genetically identified geographic ancestries and also with the observed trait s_i. The objective of admixture regression is to decompose trait variation into linear components due to genetic ancestries and linear components due to racial/ethnic group related effects.

2.2 An empirically infeasible GWAS model of ancestry-related behavioural trait variation

Admixture regression is an indirect method of analyzing group-related trait variation. In this subsection we provide a foundation for admixture regression be considering a more direct, but empirically infeasible, alternative approach based on a linear polygenic index model. Then in the next subsection we will show that the admixture regression model can be viewed as a statistically feasible simplification of this linear polygenic index model, in which proportional ancestries serve as statistical proxies for ancestry-related genetic differences.

The human genetic code contains a very large number of genetic variants (the allelles on the genome which vary between individuals) called single nucleotide polymorphisms or SNPs. Consider hypothetically a complete list of all genetic variants with any impact on variation in the observed trait. Assign a value of $0, 1$ or 2 to each SNP for individual i depending upon the number of minor allelles for that SNP. Let SNP_{iz} $i = 1, n$; denote the number of minor allelles on the zth SNP of the ith individual in the sample. The biochemical process linking human genetic variation to behavioural trait variation is unimaginably complex, and scientific understanding of the full biochemical process is very limited. Genome-wide association studies (GWAS) have made slow but steady progress in statistically modeling these linkages, although precise biochemical linkages are beyond the contemporary scientific frontier for most behavioural traits. A standard, admittedly highly simplified, model of the gene variation - trait variation nexus is the linear polygenic index model, in which the genetic component of a trait is a simple linear function of a relevant subset of the individual's genetic variants. The linear polygenic index model has been applied to a wide range of medical and behavioural traits including body mass index (Yengo et al. 2018), neuroticism (Nagel et al. 2018), depression susceptibility (Wray et al. 2018), suicidal ideation (Mullins et al. 2014), schizophrenia (Mistry et

al. 2018), educational attainment (Lee et al. 2018), neuropsychological test performance (Savage et al. 2018), and risk-taking (Clifton et al. 2017). The linear admixture regression model can be derived elegantly by invoking this standard linear polygenic index model, and hence we impose it in our model:

$$p_i = c_1 + \sum_z \beta_z SNP_{iz};\ i = 1, n, \tag{1}$$

where p_i denotes the "genetic potential" of individual i regarding the observable trait.

Let $\Pr_z(\cdot)$ denote the univariate probability distribution for SNP_z (probabilities of the three possible values $0, 1$, and 2). The probability distributions of many SNPs differ substantially across geographic ancestries, hence we define the conditional probability distributions: let $\Pr_z(\cdot|A_h = 1)$ denote the conditional probability distributions for SNP_z for individuals with purebred (that is, $A_h = 1$) ancestry for $h = 1, m$. There is no assumption of genetic homogeneity within the ancestral populations, only that they are genetically distinct and hence these (unobserved) conditional probability distributions exist as hypothetical entities. There is no attempt to estimate these conditional probability distributions directly, but rather only to use them create conditional expectations in the construction of statistical proxy variables.

The expectation of $\sum_z \beta_z SNP_{iz}$ using each purebred probability distribution defines the average genetic trait potential of each purebred ancestry:

$$\overline{p}_h = c_1 + \sum_z \beta_z E[SNP_z|A_h = 1],\ h = 1, m \tag{2}$$

which are not observed directly, but will be inferred indirectly from the admixture regression findings.

A key assumption of the admixture regression model is that admixture arises from recent random mating between the previously geographically-isolated ancestral groups. Assuming recent random mating between ancestral lines, it follows from the fundamental processes of sexual reproduction that the univariate probability distribution of any SNP for an admixed individual is the convex combination of the purebred probability distributions, with linear coefficients equal to the individual's admixture proportion. (The relationship between the multivariate distributions is more complicated, but

the multivariate distributions do not impact the expected trait given the linear polygenic index assumption.) We use a subscript . to denote the vector created from the ith row of a matrix. We assume that mating across geographic ancestries is recent and random, and therefore in particular that the univariate frequency distribution of each SNP for any individual is the convex combination of the purebred frequency distributions:

$$\Pr_z(\bullet|A_{i\cdot}) = \sum_{h=1}^{k} \Pr_z(\bullet|A_h = 1)A_{ih} \tag{3}$$

Equation (3) is a fundamental condition for the admixture regression methodology.

The linearity of genetic potential in the SNPs (1) and the random mixing assumption (3) imply that conditional expected genetic potential of an admixed individual is a convex combination of the individual's admixture proportions. Combining (1), (2) and (3) the conditional expected value of genetic potential for an individual with admixture proportions $A_{i\cdot}$ is the convex combination of the unobserved values $\overline{p}_h$ with observed linear coefficients A_{ih}:

$$E[p_i|A_{i\cdot}] = c_1 + \sum_{h=1}^{k} A_{ih}\overline{p}_h. \tag{4}$$

2.3 Ancestry proportions as a statistical proxy for ancestry-linked trait variation

A key difference in the admixture regression methodology compared to GWAS is that there is no attempt to estimate (1) directly. Rather, admixture regression uses the natural experiment of subpopulation mixing to infer differences in the conditional expected value of (1) arising from differences in the probability distribution of genetic variants across ancestries.

For expositional simplicity, in this subsection we assume that every individual included in the sample has self-identified as belonging to exactly one from the pre-specified set of k racial or ethnic groups, so that $\sum_{j=1}^{k} SIRE_{ij} = 1$ for all i. In this case, the $n \times k$ matrix of racial/ethnic group explanatory variables used in the admixture regression, denoted G, is simply set equal to

the $SIRE$ matrix: $G_{ij} = SIRE_{ij}$ for $i = 1, n; j = 1, k$. Multi-racial individuals (those who have self-identified as belonging to two or more groups) will be introduced into the analysis in the next subsection.

Define the environmental component of the trait, e_i, as the observed trait minus genetic potential:

$$s_i = p_i + e_i, \tag{5}$$

where e_i is defined as all trait variation not captured by p_i. Equation (5) is only definitional; later we will impose various conditions on e_i to enable statistical identification of the model. Define $\widetilde{p}_i$ as the genetic component of the trait for each i which is not explained by geographic ancestry $A_{i\cdot}$:

$$\widetilde{p}_i = p_i - E[p_i | A_{i\cdot}],$$

by simple substitution into (5) this gives:

$$s_i = c_1 + \sum_{h=1}^{k} A_{ih}\overline{p}_h + \widetilde{p}_i + e_i. \tag{6}$$

Recall that $\sum_{h=1}^{k} A_{hi} = 1$, for all i, so that one term in (6) is redundant; substitute $A_{i1} = 1 - \sum_{h=2}^{k} A_{ih}$ into (6) to get:

$$s_i = c_2 + \sum_{h=2}^{k} b_{Ah} A_{ih} + \widetilde{p}_i + e_i, \tag{7}$$

where $b_{Ah} = \overline{p}_h - \overline{p}_1$; $h = 2, m$, and $c_2 = c_1 + \overline{p}_1$.

Equation (7) is not well-specified as a regression model since the error term $\widetilde{p}_i + e_i$ will not be mean zero conditional on $A_{i\cdot}$ due to racial and ethnic group-related effects in e_i. In order to transform (7) into a regression model it is necessary to add explanatory terms to the regression model to remove the expected value of e_i conditional on $A_{i\cdot}$. This is accomplished by assuming that the expected differences in e_i conditional on $A_{i\cdot}$ are linearly dependent on the group identifiers $G_{i\cdot}$ and not otherwise dependent upon admixture proportions:

$$e_i = c_3 + \sum_{h=2}^{m} b_{Gh} G_{ih} + \widetilde{e}_i, \tag{8}$$

where b_{Gh} captures the environmental component associated with membership in group h relative to the reference group $h = 1$, and $\widetilde{e}_i$ is assumed to be independent of $A_{i\cdot}$, $G_{i\cdot}$ and $\widetilde{p}_i$. Although not strictly necessary, we also assume for simplicity that both residuals are normally distributed: $\widetilde{p}_i \sim N(0, \sigma_p^2)$ and $\widetilde{e}_i \sim N(0, \sigma_e^2)$. Combining (7) and (8) produces the key linear admixture regression specification:

$$s_i = c_4 + \sum_{j=2}^{k} b_{Gj} G_{ij} + \sum_{h=2}^{m} b_{Ah} A_{ih} + \varepsilon_i. \tag{9}$$

where $\varepsilon_i = \widetilde{e}_i + \widetilde{p}_i$. Note that ε_i is normally distributed with zero mean and variance $\sigma_{\widetilde{e}}^2 + \sigma_{\widetilde{p}}^2$ and is independent of $A_{i\cdot}$ and $G_{i\cdot}$.

The ordinary least squares coefficient estimates of b_{Gj}, b_{Ah}, $j = 2, k; h = 2, m$ in (9) are maximum likelihood and efficient. In many applications, the analyst also has information on the sampling substructure of the data, such as its division into site-specific subsamples. In this case, a linear mixed effects model can be used for estimating (9) rather than ordinary least squares. This involves partially decomposing the residual term ε_i in (9) into linear random effects components linked to data collection site identifiers and/or other subsample identifiers, see Heeringa and Berglund (2021).

2.4 Adding multi-racial individuals to the regression

An identifying assumption of the admixture regression technique is that the environmental influences associated with racial/ethnic group membership are captured by the group membership self-identification choices, $SIRE$. Many individuals self-identify as belonging to two or more racial or ethnic groups and the model must be adapted to this reality. In the context of our statistical framework, there are essentially three approaches: evenly splitting the individual's affiliation across their chosen groups, creating a new group for one or more particular multi-racial combinations, or deleting particular multi-racial observations where neither of the other two approaches seem appropriate.

Recall that $SIRE$ is the $n \times k$ matrix of race/ethnicity self-identifications, and we now allow that some individuals choose more than one category, so that $\sum_{j=1}^{k} SIRE_{ij} > 1$ for some i. The simplest regression specification in

this case is to assume that the group environment faced by a multi-racial individual is the average of the component group environments:

$$G_{ij} = SIRE_{ij}/(\sum_{j^*=1}^{k} SIRE_{ij^*}) \text{ for all } i = 1, n; j = 1, k. \tag{10}$$

Although (10) is a reasonable specification, it is restrictive. It is possible to replace (10) with a more general specification at some loss of parsimony. Suppose that we are concerned about imposing the restrictive condition (10) for some common multi-racial choice (such as, for example, Black-White biracial in a US dataset). Let V_1 denote a $k-$vector with ones for the included race/ethnicity groups in this particular multi-racial combination and zeros elsewhere. We can supplement (10) by adding a $k+1^{st}$ group and using a different rule for this subset of multi-racials:

$$\begin{aligned} G_{ij} &= 0 \text{ for } j = 1, k \text{ if } SIRE_{i\cdot} = V_1 \\ G_{i,k+1} &= 1 \text{ if } SIRE_{i\cdot} = V_1 \\ G_{i,k+1} &= 0 \text{ if } SIRE_{i\cdot} \neq V_1, \end{aligned} \tag{11}$$

where $SIRE_{i\cdot} = V_1$ denotes vector equality between these two $k-$vectors. There are now $k+1$ groups: the originally specified SIRE groups and a new group for the selected multiracial combination. G becomes a $n \times (k+1)$ matrix, and the regression (9) described in the previous subsection applies exactly as before but with one extra dimension to G. Any small number of defined multi-racial groups can be appended in this way. The only change to the regression methodology is that G becomes a $n \times k^*$ matrix (with an associated increase in the set of estimated parameters) where $k^* - k$ is the number of multiracial combinations added as new categories.

It is not feasible to use rule (11) for all race/ethnicity choice combinations due to lack of parsimony; there are $2^k - k$ potential multi-racial combinations and each one added requires an additional parameter in the regression. It can only be used for the common multi-racial choices where there is sufficient data of that combination in the sample. For all others, it is necessary to stick with the restrictive assumption (10) or drop the observations from the sample. This will be illustrated in the empirical application in Section 4.

Once a regression model is estimated using (11), it is possible to test the accuracy of restrictive assumption (10) for that multi-racial group. The

restrictive assumption implicit in (10) requires that the average of the coefficients of the components equals the added-group coefficient in the unrestricted model:

$$\frac{1}{\#j^*}\sum_{j^*} b_{Gj^*} = b_{G,k+1}, \tag{12}$$

where $\#j^*$ denotes the number of components in the multiracial category (typically either two or three) and the sum runs over these element only. This is a linear restriction on the vector of coefficients, or multiple linear restrictions for $k^* - k$ greater than one, which can be tested with a t-test (for each group coefficient singly) or a Wald test for all them, as detailed below.

Let $\widehat{b}$ denote the $m + k^*-$vector of all the coefficients in the admixture regression (9):

$$\widehat{b} = [\widehat{c}_4, \widehat{b}_G, \widehat{b}_A],$$

and let $\widehat{Cov}_{\widehat{b}}$ denote the estimated $(m + k^*) \times (m + k^*)$ −covariance matrix of these estimates.

First consider the case $k^* - k = 1$. Let R denote the $(m + k^*)-$vector expressing restriction (12) imposed on b. For example, if the group combination consists of individuals who choose all three of the first, second, and third SIRE categories (recalling that the first SIRE category has a zero beta by definition) the restriction vector is:

$$R = [0, -\frac{1}{3}, -\frac{1}{3}, 0, .., 0, 1, 0, ..., 0]$$

where the 1 is element k^* in the vector. Any other restriction of type (12) is easily stated in this way. In the case of one group, this gives rise to a standard t-test of the one coefficient restriction, and in particular:

$$\frac{\widehat{b}'R}{(R'\widehat{Cov}_{\widehat{b}}R)} \sim t(n - m - k^*). \tag{13}$$

For the case $k^* - k > 1$ it is possible to test each multi-racial group equality individually as above using (13) or perform a joint Wald test on all of them. Let R denote the $(m + k^*) \times (k^* - k)-$matrix of all the linear restrictions, giving the standard Wald test:

$$\widehat{b}'R(R'(\widehat{Cov}_{\widehat{b}})^{-1}R)^{-1}R'\widehat{b} \overset{a}{\sim} \chi^2(k^* - k) \tag{14}$$

where $\overset{a}{\sim}$ denotes the approximate distribution for large n. In the case of estimation by linear mixed effects modeling, both test statistics (13) and (14) are large$-n$ asymptotic distributions rather than exact finite-sample distributions, but they remain valid tests.

3 Extensions of the linear admixture regression model

3.1 Additional explanatory variables with and without orthogonalization

It is straightforward to include additional explanatory variables in the admixture regression model. Let $x_{i1}, x_{i2}, ..., x_{il}$ denote a set of explanatory variables that help to linearly explain the trait along with the ancestry proportions and group identities. We modify specification (9) to include these:

$$s_i = +\sum_{j=2}^{k} b_{Gj} G_{ij} + \sum_{h=2}^{m} b_{Ah} A_{ih} + \sum_{d=1}^{l} b_{xd} x_{id} + \varepsilon_i \tag{15}$$

and keep all the other assumptions as before. The estimation theory for (15) is essentially identical to that of (9) as discussed above.

In some cases, the admixture regression model with additional explanatory variables (15) can be made more useful and informative by orthogonal rotation of one or more of the explanatory variables, in order to aggregate the full linear effects of proportional ancestries and group identities into their associated coefficients. To understand why such an orthogonal rotation might be useful, consider the hypothetical case of an admixture regression model of Body Mass Index (BMI) in which waist measurement is one of the explanatory variables. Waist measurement has such strong explanatory power for BMI that its presence in an admixture regression model like (15) will diminish the direct explanatory power of proportional ancestries and group identities; their total impact will be partly hidden within the waist measurement variable. This can be remedied by orthogonalizing the waist measurement variable with respect to the proportional ancestry and group identity variables before estimating the admixture regression, as explained next.

Suppose that variable x_1 in (15) has strong explanatory power for s and substantial correlation with proportional ancestry and/or group identity vari-

ables, and therefore the analyst wishes to orthogonalize it with respect to G_{ij} and A_{ih}, $j = 2, k; h = 2, m$. In a first step, the analyst can perform a simple least square regression decomposition of x_1 into the component linearly explained by these variables, and the residual, orthogonal component x_1^o:

$$x_1 = \widehat{c}_7 + \sum_{j=2}^{k} \widehat{b}_{Gj} G_{ij} + \sum_{h=2}^{m} \widehat{b}_{Ah} A_{ih} + x_1^o \tag{16}$$

Since all the explanatory variables are deterministic (that is, conditionally fixed variables rather than random variables in the regression model), this orthogonalization step (16) is interpreted as a matrix transformation of fixed vectors and does not alter any statistical assumptions of the main regression model. It merely serves to linearly rotate the deterministic explanatory variables used in the actual, second-stage, admixture regression. Replacing x_1 with x_1^o in (15) changes the interpretation of the coefficients $\widehat{b}_{Gj}$ and $\widehat{b}_{Ah}$, $j = 2, k; h = 2, m$ since they now include the G_{ij} and A_{ih} related explanatory power from x_1. An illustrative example will be provided in Section 4 below.

3.2 A semiparametric extension of the admixture regression model

The linear dependence of the trait on admixture proportions in our regression model is in part an artifact of the assumption of a linear polygenic index (1). It is possible to weaken this linearity assumption using nonparametric regression methods. We replace the restrictive assumption of a linear polygenic index (1) with a very general description of genetic potential as a function of the full vector of genetic variants:

$$p_i = p(SNP_{i\cdot})$$

and instead of linearity as in (1) only require smoothness conditions on the conditional expectation of $p(\cdot)$ as a function of the ancestral proportions vector, as delineated below.

As in earlier subsections, we consider p_i as a stochastic function of the ancestral proportions vector $A_{i\cdot}$, but now without imposing the strict linearity (4) arising from the linear polygenic index assumption:

$$f(A_{i\cdot}) = E[p(SNP_{i\cdot})|A_{i\cdot}].$$

Define the unexplained component of p_i as before:

$$\widetilde{p}_i = p_i - f(A_{i\cdot})$$

and as before we assume that $\widetilde{p}_i \sim N(0, \sigma^2_{\widetilde{p}})$ and independent of $A_{i\cdot}$ and $G_{i\cdot}$. We impose exactly the same assumptions on e_i as in Subsection 2.2, giving:

$$s_i = \sum_{j=2}^{k} b_{Gj} G_{ij} + f(A_{i\cdot}) + \varepsilon_i, \tag{17}$$

where $\varepsilon_i = \widetilde{p}_i + \widetilde{e}_i$ is normally distributed with mean zero and variance $\sigma^2_{\widetilde{p}} + \sigma^2_{\widetilde{e}}$ and independent of $A_{i\cdot}$ and $G_{i\cdot}$. This equation (17) is a partially linear nonparametric regression model, see, e.g. Li and Racine (2007). This model can be consistently estimated using the three-step procedure of Robinson (1988). We will impose Condition 7.1 from Li and Racine (2007) in order to justify this procedure within our framework (see the Technical Appendix for details).

For the case $m > 2$ the general specification (17) suffers from the curse of dimensionality and is unlikely to be estimable on moderate-sized datasets. A more restrictive specification is needed to give the model sufficient parsimony for estimation. One reasonable specification choice is to restrict the nonlinearity in the impact of ancestries on the trait to a single ancestral category, which we assume is ancestry category 2, giving rise to the specification:

$$s_i = \sum_{j=2}^{k} b_{Gj} G_{ij} + f_2(A_{i2}) + \sum_{h=3}^{m} b_{Ah} A_{ih} + \varepsilon_i, \tag{18}$$

and we will now rely on this more restrictive specification throughout the remainder of this subsection.

We assume that the unconditional density $\Pr(A_2)$ is continuous and strictly positive everywhere on the $[0, 1]$ interval. Let $\widehat{\Pr}(A_{i2})$ denote the nonparametrically estimated unconditional density of A_{i2} :

$$\widehat{\Pr}(A_{i2}) = \frac{1}{n} \sum_{i'=1}^{n} k(A_{i'2} - A_{i2}), \tag{19}$$

where $k(\bullet)$ is a kernel weighting function. In our empirical application in Section 4 we use the Gaussian kernel weighting function, $k(x) = \frac{1}{\phi\sqrt{2\pi}} e^{-(x/\phi)^2}$ where ϕ is the chosen bandwidth.

In the first step of the Robinson procedure, the conditional means of the dependent variable and linear-component explanatory variables are estimated nonparametrically as functions of the nonparametric-component explanatory variable, A_{i2}:

$$\widehat{f}_0(A_{i2}) \approx E[s_i|A_{i2}]$$

$$\widehat{f}_{Gj}(A_{i2}) \approx E[G_{ij}|A_{i2}];\ j = 2, k$$

and

$$\widehat{f}_{Ah}(A_{i2}) \approx E[A_{ih}|A_{i2}];\ h = 3, m$$

that is:

$$\widehat{f}_0(A_{i2}) = \frac{1}{n}\sum_{i'=1}^{n} s_{i'} k(A_{i'2} - A_{i2})/\widehat{\Pr}(A_{i2}),$$

$$\widehat{f}_{Gj}(A_{i2}) = \frac{1}{n}\sum_{i'=1}^{n} G_{i'j} k(A_{i'2} - A_{i2})/\widehat{\Pr}(A_{i2}),\ j = 2, k.$$

and

$$\widehat{f}_{Ah}(A_{i2}) = \frac{1}{n}\sum_{i'=1}^{n} G_{i'j} k(A_{i'2} - A_{i2})/\widehat{\Pr}(A_{i2}),\ 3 = 2, m.$$

In the second step, the linear parameters of the model (17) are estimated by ordinary least squares, replacing the dependent variable and linear-component explanatory variables with the deviations from their conditional mean functions:

$$(\widehat{b}_{G\cdot}, \widehat{b}_{A\cdot}) = (X'X)^{-1}X'y$$

where

$$y_i = s_i - \widehat{f}_0(A_{i2})$$

$$X_{ij} = G_{ij} - \widehat{f}_{Gj}(A_{i2});\ j = 2, k,$$

$$X_{ih} = A_{ih} - \widehat{f}_{Ah}(A_{i2});\ h = 3, m.$$

Note that $(\widehat{b}_{G\cdot}, \widehat{b}_{A\cdot})$ is a $(k+m-3)-$vector and X is a $n \times (k+m-3)-$matrix where the index first runs from 2 to k over j and then from 3 to m over h.

In the third step, the nonparametric component of the model is estimated by subtracting the predicted linear component from both sides of (17) and then applying standard nonparametric regression:

$$y_i^* = s_i - (\sum_{j=2}^{k} \widehat{b}_{Gj} G_{ij} + \sum_{h=3}^{m} \widehat{b}_{Ah} A_{ih});\ i = 1, n$$

and then:

$$\widehat{f}(A_{i2}) = \frac{1}{c_i}\sum_{i'=1}^{n} k(A_{i'2} - A_{i2})y^*_{i'},$$

where $c_i = \sum_{i'=1}^{n} k(A_{i'2} - A_{i2})$.

The partially linear nonparametric approach to admixture regression is more empirically challenging than the linear specification. Proper implementation of the technique involves a tradeoff between parsimony, the generality of the specification used, and the distributional features of the available data. An example of (18) will be estimated in Section 4 below.

4 Empirical Application

In this section, we illustrate the techniques by performing an admixture regression analysis of neuropsychological performance from the Adolescent Brain Cognitive Development (ABCD) database. The ABCD study is the largest long-term study of brain development and child health in the United States, testing 11,000 children ages 9-10 at 21 testing sites; see Karcher and Barch (2020) for an overview. Our sample consists of age and gender-adjusted scores and genotyped DNA samples of the 9972 children who met our sample selection criteria, along with questionaire responses of their parent(s)/guardian(s). The dependent variable in our model is the composite neuropsychological performance score based on the NIH Toolbox® (NIHTBX) neurocognitive battery provided in the ABCD database; this consists of tasks measuring attention, episodic memory, language abilities, executive function, processing speed, and working memory. Our core explanatory variables are seven SIRE variables, White, Black, Hispanic, Native American, East Asian, South Asian, and Other (and including multiple SIRE choices from among these) and five genetic ancestry proportions of European, African, Amerindian, East Asian and South Asian background obtained from the genotyped DNA samples. Children whose parent(s)/guardian(s) identified the child as belonging to Pacific Islander racial groups were excluded from our analyses owing to a lack of corresponding ancestry category in our chosen five categories. The ABCD Version 3 database provides 516,598 genotyped SNP variants for each individual's DNA sample. After quality control, filtering, and pruning we were left with 99,642 SNP variants to determine

the five ancestry proportions, employing the Admixture 1.3 software package (Alexander et al. 2015). See the Supplemental Materials for more detailed description of the ABCD database, our sample selection procedure, and the construction of the variables that we use.

Table 1 shows means and standard deviations of the regression dependent variable on data subsets sorted by SIRE choice. On the full sample, by construction, the dependent variable has a mean of zero and standard deviation of one. There is considerable dispersion in the subsample means sorted by SIRE; for example, the means differ by 1.02 standard deviations (using the full-sample standard deviation for simplicity) between two of the largest SIRE categories shown, White-only SIRE and Black-only SIRE. The considerable variation in means for SIRE-based subsamples provides an initial justification for performing admixture regression analysis. This is a table of descriptive statistics; the standard errors shown are not appropriate for formal hypothesis testing since there is no adjustment for potential site-linked and family-linked correlations, particularly relevant in the case of the smaller subsample categories.

Table 2 displays empirical results from three specifications of the admixture regression methodology. Recall that one SIRE variable and one ancestry proportion variable must be left out as an identification condition of the admixture regression: we leave out the White SIRE variable and the European ancestry proportion variable. Model 1 uses a linear regression specification and singleton SIRE categories for the group-identity variables G; individuals who choose multiple SIRE categories have G exposures equally divided between the chosen SIRE categories as in (10). Three of the four ancestral proportion variables and one of the six group-identity variables have statistically significant coefficients. Model 2 adds a selected set of multiple-SIRE composite categories to the G specification. We include the seven two-category choices with the largest number of observations in our sample. Individuals with one of these two-category choices has unit exposure to the associated explanatory variable, and no exposure to the weighted single-SIRE variables (see equation 11 above). The same three of four ancestral proportion variables as in Model 1 are significant in Model 2, with similar coefficients to Model 1. None of the single-SIRE group identity variables is significant. Five of the seven selected two-SIRE group identity variables have significantly different coefficients from that implied by equal weightings of the component single-category coefficients. One of these (Hispanic-Other) has a statistically significant coefficient; the other four are not significantly

different from zero, but are significantly different from the value implied by the composite single-category coefficients. Random effects are included in all models except Model 3 to capture any common variation associated with the 22 individual data collection sites in the ABCD study or associated with those families having multiple individuals in the sample. We use the *lmer* maximum likelihood mixed effects model estimation routine from the *R* language library, see Bates et al. (2015), for all models except Model 3. See Nakagawa and Schielzeth (2013) for the definition and interpretation of conditional and marginal R^2 in a linear mixed effects model.

Model 3 implements a partially linear nonparametric specification. This specification requires that the highlighted ancestry proportion (whose impact is estimated nonparametically) has observations throughout the [0,1] range. For each of the five ancestry categories, Table 3 gives the number of sample observations of proportional ancestry in decile bins of percent ancestry, for each of the five genetic ancestry categories. We use African proportional ancestry as the highlighted variable since it fulfils the requirement for observations throughout the $[0, 1]$ interval and therefore partially linear nonparametric estimation is feasible. Figure 1 shows the probability density of African ancestry for the full sample population; Figure 2 shows the density restricted to those individuals having measured African ancestry greater than 0.5%, this provides greater detail in the graph by excluding observations with near-zero ancestry. Interestingly, this density has three local peaks, at approximately 5%, 40% and 80% African ancestry.

Partially linear semiparametric Model 3 (18) is estimated using the *npplr* routine in the R programming language subroutine library *NP* written and maintained by Hayfield and Racine (2020). We use the simple average SIRE specification of G as in Model 1. We use the Guassian kernel throughout, and all bandwidths are chosen by iterated least-squares cross-validation. The linear coefficient estimates in Model 3 do not differ notably from those in Model 1. Figure 3 displays the nonparametric estimate of the impact of African ancestry on the performance variable along with the corresponding linear impact estimate from Model 1, that is, $\widehat{f}(A_2) - \widehat{f}(0)$ and $A_2\widehat{b}_2$ for $A_2 \in [0, 1]$. There is some graphical evidence for an uptick in the nonlinear gradient for ancestry proportions above 90%. We now briefly examine this further.

Model 3 does not capture the efficiency gain and test statistic bias reduction from the mixed effects modeling used in the estimation of the other models. Figure 3 of Model 3 is estimated in the second stage of a two-stage

semiparametric estimation process and this weakens its empirical reliability. To examine more carefully the graphical pattern observed in Figure 3, but with single-stage estimation and the advantage of mixed effects modeling, we estimate a piecewise linear specification for $A_{i2} \geq 0.9$. This was chosen in order to mimick the observed nonlinear uptick seen in Figure 3 within a linear regression functional form. Recall that African ancestry proportion is ancestry variable 2, giving the formulation:

$$s_i = c + \sum_{j=2}^{7} b_{Gj} G_{ij} + \sum_{h=2}^{5} b_{Ah} A_{ih} + b^{kink} A_{i2} D[A_{i2} \geq 0.9] + \varepsilon_i, \qquad (20)$$

where $D[\bullet]$ is a zero-one dummy variable and b^{kink} is the added coefficient. The results are shown as Models 4 and 5 in Table 4. In Model 4 we use the simple average SIRE specification of G as in Model 1; Model 5 adds the same seven two-SIRE combination groups as in Model 2. The coefficient b^{kink} is significantly positive in one of the two models; the significance of this finding must be treated with caution since the particular kink specification (20) is based on examination of Figure 3 using the same data.

Table 5 adds two new variables, US born child and Social-Economic Status (SES), to the admixture regression model. US born child equals one if the child was born in the USA and zero if born elsewhere. SES is a factor-analytic composite of underlying variables from the ABCD database including neighborhood SES, subjective SES as determined from a set of questionaire answers by the parent(s)/guardian(s) of the child on parental/guardian marital status, completed level of parental/guardian education, reported neighborhood safety, and parental/guardian employment. See the Supplemental Materials for more detailed discussion. Models 6 and 7 are identical to Models 4 and 5 (respectively) from Table 3, except for the addition of these two variables. As discussed in Section 3 above, including additional explanatory variables complicates the interpretation of an admixture regression model in terms of the implied decomposition of trait variation into linear components linked to group identities and components linked to genetic ancestries. The SES variable covaries strongly with both genetic and environmental components of neuropsychological performance scores. To retain the standard interpretability of the admixture regression it is important to orthogonalize SES with respect to the group identity and ancestry variables before running the regression. For completeness, Models 6 and 7 are shown with and without the orthogonalization of SES (versions a and b of each model). If the

purpose of the estimation is to identify the total impact of SES on the trait, the regression with raw SES is more appropriate (version a). For admixture analysis intended to capture the total effects of group identity and genetic ancestry on the trait, orthogonalized SES is more appropriate (version b).

5 Conclusion

Many behavioural traits covary strongly with racial/ethnic self-identities, but it is often ambigous whether this covariance reflects environmental causes associated with racial/ethnic identity groups or reflects underlying genetic similarity among group members arising from shared geographic ancestry. Admixture regression relies on the natural experiment of recent genetic admixture of previously geographically-isolated ancestral groups to measure the explanatory power arising from racial/ethnic group identities and that arising from ancestry-based similarities of genetic background. The admixture regression methodology, in various formulations, has been applied to a wide range of medical and behavioural traits including asthma, obesity, type 2 diabetes, hypertension, neuropsychological performance, and sleep depth.

This paper provides a statistical framework for admixture regression based on the linear polygenic index model of behavioural genetics, and develops refinements and extensions of the methodology within this framework. We provide a simple new test procedure for determining whether multiple-SIRE categories have independent explanatory power not captured by the individual component categories. We consider additional explanatory variable in the admixture regression and their interpretation with and without orthogonalization with respect to core variables. We weaken the linearity assumption and develop a partially linear semiparametric regression specification. We illustrate our methodology using neuropsychological performance test data from the Adolescent Brain Cognitive Development database, but the techniques have broader applicability.

Bibliography

Alexander, D.H., S.S. Shringarpure, J. Novembre and K. Lange (2015). Admixture 1.3 Software Manual, Simon Laboratory, University of Wisconsin, Bioinformatics Programs.

Bates, D., M. Mächler, B.M. Bolker, S.C. Walker (2015) Fitting Linear Mixed-Effects Models Using lme4, *Journal of Statistical Software*, Vol. 67: 1-48.

Cheng, C.Y., D. Reich, C.A. Haiman, A. Tandon, N. Patterson, S. Elizabeth, E.L. Akylbekova, F.L. Brancati, J. Coresh, E. Boerwinkle, D. Altshuler, H.A. Taylor, B.E. Henderson, J.G. Wilson, W.H.L. Kao (2013), African Ancestry and Its Correlation to Type 2 Diabetes in African Americans: A Genetic Admixture Analysis in Three U.S. Population Cohorts, *PLOS One*, Vol. 7: 1-9.

Choquet, H., Yin, J., and Jorgenson, E. (2021). Cigarette smoking behaviors and the importance of ethnicity and genetic ancestry. *Translational psychiatry*, Vol. 11: 1-10.

Choudhry, S., E.G. Burchard, L.N. Borrell, H. Tang, I. Gomez, M. Naqvi, et alia (2006). Ancestry–environment interactions and asthma risk among Puerto Ricans. *American Journal of Respiratory and Critical Care Medicine* Vol. 174: 1088-1093.

Clifton, E.A.D., J.R.B. Perry, F. Imamura, F.R. Day, et alia. (2018) Genome–wide association study for risk taking propensity indicates shared pathways with body mass index, *Communications Biology*, 1-36.

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

Hayfield, T. and J.S. Racine (2020). Nonparametric kernel smoothing methods for mixed data types. R package np documentation.

Heeringa, S.G. and P.A. Berglund (2020). A guide for population-based analysis of the Adolescent Brain Cognitive Development (ABCD) Study baseline data. BioRxiv preprint.

Karcher, N.R. and D.M. Barch (2021). The ABCD study: understanding the development of risk for mental and physical health outcomes. *Neuropsychopharmacology* Vol. 46: 131–142.

Klimentidis, Y.C., G.F. Miller and M.D. Shriver (2009). The relationship between European genetic admixture and body composition among Hispanics and Native Americans, *American Journal of Human Biology* Vol. 21: 377-82.

Klimentidis, Y.C., A. Dulin-Keita, K Casazza, A.L. Willig, D.B. Allison and J.R. Fernandez (2012). Genetic admixture, social–behavioural factors and body composition are associated with blood pressure differently by racial–ethnic group among children, *Journal of Human Hypertension* Vol. 26: 98–107.

Lackland, D.T. (2014). Racial differences in hypertension: implications for high blood pressure management. *The American Journal of the Medical Sciences*, Vol. 348: 135-138.

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

Lee, J.J., R. Wedow, A. Okbay, et alia (2018). Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nature Genetics,* Vol. 50: 1112–1121.

Li, Q. and J.S. Racine (2007) *Nonparametric econometrics: theory and practice*, Princeton University Press, Princeton NJ, USA.

Llibre-Guerra, J.J., Y. Li, I.E. Allen, J.C. Llibre-Guerra, A.M. Rodríguez Salgado, A.I. Peñalver, et alia (2021). Race, genetic admixture and cognitive performance in the Cuban population. Forthcoming in *The Journals of Gerontology: Series A*.

Mistry, S., Harrison, J. R., Smith, D. J. , Escott-Price, V. and Zammit, S. (2018) The use of polygenic risk scores to identify phenotypes associated with genetic risk of schizophrenia: systematic review. *Schizophrenia Research*, Vol. 197: 2-8.

Mullins, N., et alia (2014). Genetic relationships between suicide attempts, suicidal ideation and major psychiatric disorders: a genome-wide association and polygenic scoring study. American Journal of Medical Genetics. Part B, *Neuropsychiatric Genetics: the Official Publication of the International Society of Psychiatric Genetics* Vol. 165B: 428-437.

Nagel, M., 23andMe Research Team, et alia (2018). Meta-analysis of genome-wide association studies for neuroticism in 449,484 individuals identifies novel genetic loci and pathways. *Nature Genetics,* Vol. 50: 920–927.

Nakagawa, S. and H. Schielzeth (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models, *Methods in Ecology and Evolution*, Vol. 4: 133-142.

Robinson, P.M. (1988). Root-n consistent semiparametric regression. *Econometrica, Vol.* 56: 931-954.

Ruderman, A., L.O. Pérez, K. Adhikari, P. Navarro, V. Ramallo, C. Gallo, R. González-José, et alia (2019). Obesity, genomic ancestry, and socioeco-

nomic variables in Latin American mestizos. *American Journal of Human Biology*, Vol. 31: 1-13.

Salari K., S. Choudhry, H. Tang, et alia (2005). Genetic admixture and asthma-related phenotypes in Mexican American and Puerto Rican asthmatics. *Genetic Epidemiology* Vol. 29: 76–86.

Savage, J.E., et alia. (2018). Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. *Nature Genetics,* Vol. 50: 912–919.

Wang, Y., and M.A. Beydoun (2007). The obesity epidemic in the United States—gender, age, socioeconomic, racial/ethnic, and geographic characteristics: a systematic review and meta-regression analysis. *Epidemiologic reviews*, Vol. 29: 6-28.

Williams, R.C., J.C. Long, R.L. Hanson, M.L. Sievers, and W.C. Knowler (2000). Individual Estimates of European Genetic Admixture Associated with Lower Body-Mass Index, Plasma Glucose, and Prevalence of Type 2 Diabetes in Pima Indians, *American Journal of Human Genetics*, Vol. 66: 527–538

N. R. Wray et al.; eQTLGen; 23andMe; Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium (2018). Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nature Genetics,* Vol. 50: 668–681.

L. Yengo et al.; GIANT Consortium (2018). Meta-analysis of genome-wide association studies for height and body mass index in ~700000 individuals of European ancestry. *Human Molecular Genetics.* Vol. 27: 3641–3649.

Technical Appendix

In this technical appendix we re-state condition 7.1 from (Racine and Li 2007, p. 224) in the context of our partially linear admixture regression model (18).

We assume that the $(k+m-1)-$vector of observations (s_i, G_{ij}, A_{ih}) $j = 2, k; h = 2, m$ has an i.i.d. distribution over observations $i = 1, n$ and that the conditional mean functions $E[G_{ij}|A_{i2}]$ and $E[A_{ih}|A_{i2}]$ are twice differentiable throughout the interior of the domain of A_2, the closed unit interval . Let $m(\bullet)$ denote any of these conditional mean functions or their first or second derivative functions. As in Racine and Li, we impose the following Lipschitz-type smoothness condition on these conditional mean functions and their first and second derivatives: $|m(A_2) - m(A_2')| \leq H(z)|A_2 - A_2'|$ where $H(\bullet)$ is some continuous function such that $E[H(A_2)^2]$ is finite. The expectation of $H(A_2)^2$ is over the probability distribution of A_2.

We continue to assume that ε_i is mean-zero normally distributed with constant variance. Since G_{ij} only takes the values of zero and one and A_{ih} is confined to the unit interval, it necessarily follows that both have bounded fourth moments. We assume that $k(\bullet)$ is a bounded second-order kernel.

To formally derive the limiting distribution of the Robinson estimator, it is necessary to define a trimming parameter which ensures that the estimates $\widehat{\Pr}(A_{i2})$ are bounded away from zero. Let t denote a trimming parameter and consider the estimator described in the text but where observations such that $\widehat{\Pr}(A_{i2}) < t$ in (19) are dropped from the subsequent estimation steps. Let ϕ denote the kernel bandwidth for sample size n. Assume that the trimming parameter obeys the following two limiting conditions as $n \to \infty$: $n\phi^2 t^4 \to \infty$ and $nt^{-4}\phi^8 \to 0$.

Under these conditions we have from (Robinson 1988) that

$$d \lim \sqrt{n}[(\widehat{b}_{G\cdot}, \widehat{b}_{A\cdot}) - (b_{G\cdot}, b_{A\cdot})] \sim N(0, \sigma_\varepsilon^2 E[X'X]^{-1}).$$

where the matrix X is defined in the main text of the paper above, in step two of the Robinson procedure.

Table 1

Neuropsychological Performance Scores Sorted by Self-identified Race or Ethnicity (SIRE)

Within-Category Means and Standard Deviations	Individuals with Single-SIRE Identities						All Hispanic
	White Only	Black Only	Native American Only	East Asian Only	South Asian Only	Other Only	
Mean	0.25	-0.77	-0.42	0.57	0.45	-.22	-.23
Standard Deviation	0.92	0.87	0.79	1.02	1.02	1.12	0.96
Number of Observations	5593	1434	31	107	43	97	1869
	Individuals with Selected Multiple-SIRE Identities						
	Black-White	Hispanic-White	Native American - White	East Asian - White	South Asian - White	Hispanic - Black	Hispanic - Other
Mean	-0.13	-0.19	0.01	0.58	0.83	-0.34	-0.45
Standard Deviation	0.95	0.96	0.86	0.98	0.84	0.93	0.92
Number of Observations	302	1171	131	249	40	84	411

Notes to Table: The table shows means and standard deviations of neuropsychological performance scores for individuals sorted into categories by self-identified race and ethnicity (SIRE). The categories used are: individuals who choose only one of the six race categories (White, Black, Native American, East Asian, South Asian, and Other), all individuals selecting Hispanic ethnicity along with any of the six race categories, and individuals choosing the seven most common two-SIRE choices. By construction the mean and standard deviation of the full sample are zero and one; there are 9972 individuals in the full sample.

Table 2

Admixture Regression Results for Neuropsychological Performance

Linear Specifications with and without Composite Groups and a Partially Linear Semiparametric Specification

	Core Explanatory Variables										
	Intercept	% African	% Amerindian	% East Asian	% South Asian	Black SIRE	Hispanic SIRE	Native American SIRE	East Asian SIRE	South Asian SIRE	Other SIRE
Model 1	0.3010	-1.0242	-1.3804	0.6439	0.4867	-0.1420	-0.0940	-0.1428	-0.2120	-0.0735	-0.1967
t-statistic	9.0640	-8.3380	-11.3700	3.1850	1.4910	-1.4050	-1.1150	-1.3340	-1.2060	-0.2820	-2.5210
Model 2	0.2949	-1.0298	-1.3358	0.6602	0.5439	-0.1503	0.3390	-0.1856	-0.3523	-0.2679	-0.1796
t-statistic	8.9330	-8.2250	-10.7830	3.2530	1.6590	-1.4620	1.7250	-1.3720	-1.9280	-0.9930	-1.7640
Model 3	N/A	Figure 3	-1.1914	0.6924	0.7377	-0.1289	-0.1202	-0.2578	-0.1369	-0.1523	-0.0761
t-statistic			-9.9517	3.4772	2.3109	-1.3291	-1.4464	-2.5228	-0.7736	-0.5955	-0.9088
	Multiple-SIRE-Composite Explanatory Variables										
	Black-White SIRE	Hispanic-White SIRE	Native America - White SIRE	East Asian - White SIRE	South Asian - White SIRE	Hispanic - Black SIRE	Hispanic - Other SIRE				
Model 2 [cont.]	0.0533	-0.0813	-0.0999	0.0043	0.3361	0.0361	-0.1657	Wald Test Statistic	Wald Test p-value		
t-statistic	0.7110	-1.7890	-1.1850	0.0420	1.8000	0.2940	-2.4000				
Test 2	2.2813	-2.5535	-0.0689	2.4659	2.8140	-1.6174	-2.1221	26.4231	0.0004		
Conditional R2	Model 1: 0.550; Model 2: 0.550; Model 3:NA										
Marginal R2	Model 1: 0.157; Model 2: 0.160; Model 3:NA										

Notes to Table: Model 1 uses single-SIRE categories with multiple-SIRE choices allocated evenly across them; Model 2 adds seven multiple-SIRE categories. Model 4 uses semiparametric estimation and single-SIRE categories as in Model 1. Test 2 gives the z-statistic for testing if the multiple-SIRE group coefficient equals the average of the component coefficients; the Wald statistic provides a joint test of all the Test 2 restrictions.

Table 3

Number of Observations in Deciles of Proportional Ancestry for Each Ancestry Category

European	Interval	$A_{1i} \leq 10\%$	$10\% < A_{1i} \leq 20\%$	$20\% < A_{1i} \leq 30\%$	$30\% < A_{1i} \leq 40\%$	$40\% < A_{1i} \leq 50\%$
	Number of Obs.	298	908	425	346	461
	Interval	$50\% < A_{1i} \leq 60\%$	$60\% < A_{1i} \leq 70\%$	$70\% < A_{1i} \leq 80\%$	$80\% < A_{1i} \leq 90\%$	$90\% < A_{1i}$
	Number of Obs.	700	406	462	514	5452
African	Interval	$A_{2i} \leq 10\%$	$10\% < A_{2i} \leq 20\%$	$20\% < A_{2i} \leq 30\%$	$30\% < A_{2i} \leq 40\%$	$40\% < A_{2i} \leq 50\%$
	Number of Obs.	7557	2935	286	125	165
	Interval	$50\% < A_{2i} \leq 60\%$	$60\% < A_{2i} \leq 70\%$	$70\% < A_{2i} \leq 80\%$	$80\% < A_{2i} \leq 90\%$	$90\% < A_{2i}$
	Number of Obs.	88	130	406	787	149
Amerindian	Interval	$A_{3i} \leq 10\%$	$10\% < A_{3i} \leq 20\%$	$20\% < A_{3i} \leq 30\%$	$30\% < A_{3i} \leq 40\%$	$40\% < A_{3i} \leq 50\%$
	Number of Obs.	8364	443	329	301	282
	Interval	$50\% < A_{3i} \leq 60\%$	$60\% < A_{3i} \leq 70\%$	$70\% < A_{3i} \leq 80\%$	$80\% < A_{3i} \leq 90\%$	$90\% < A_{3i}$
	Number of Obs.	156	75	18	0	4
East Asian	Interval	$A_{4i} \leq 10\%$	$10\% < A_{4i} \leq 20\%$	$20\% < A_{4i} \leq 30\%$	$30\% < A_{4i} \leq 40\%$	$40\% < A_{4i} \leq 50\%$
	Number of Obs.	9455	74	84	20	225
	Interval	$50\% < A_{4i} \leq 60\%$	$60\% < A_{4i} \leq 70\%$	$70\% < A_{4i} \leq 80\%$	$80\% < A_{4i} \leq 90\%$	$90\% < A_{4i}$
	Number of Obs.	18	4	8	10	74
South Asian	Interval	$A_{5i} \leq 10\%$	$10\% < A_{5i} \leq 20\%$	$20\% < A_{5i} \leq 30\%$	$30\% < A_{5i} \leq 40\%$	$40\% < A_{5i} \leq 50\%$
	Number of Obs.	9796	55	30	29	17
	Interval	$50\% < A_{5i} \leq 60\%$	$60\% < A_{5i} \leq 70\%$	$70\% < A_{5i} \leq 80\%$	$80\% < A_{5i} \leq 90\%$	$90\% < A_{5i}$
	Number of Obs.	4	11	9	21	0

Notes to Table: For each of the five geographic ancestries, the table shows the number of the total 9972 observations within each of the deciles of proportional ancestry.

Figure 1

Estimated Density of African Ancestry for the Full Sample

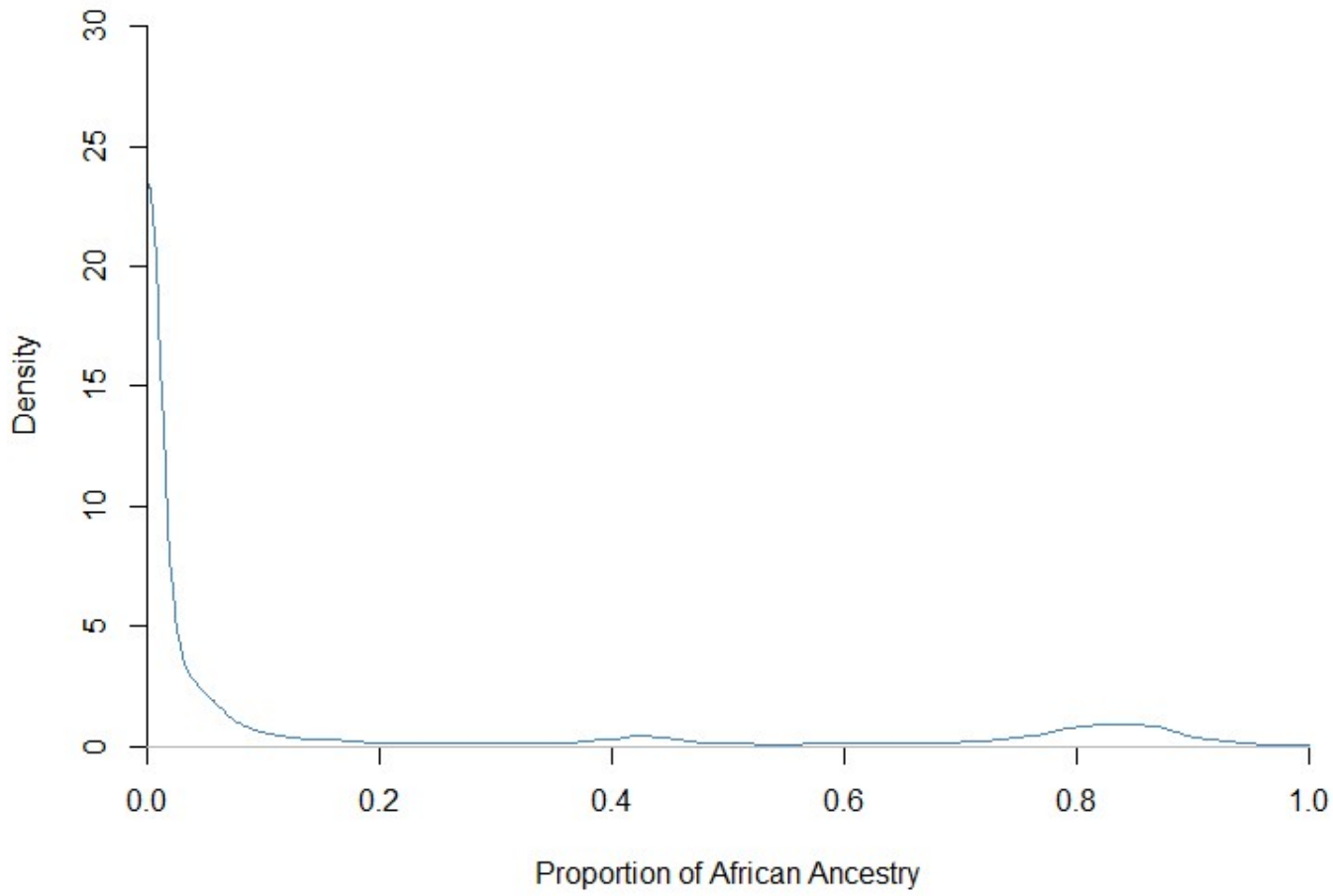

Figure 2

Estimated Density of African Ancestry for a Restricted Sample (Ancestry > 0.5%)

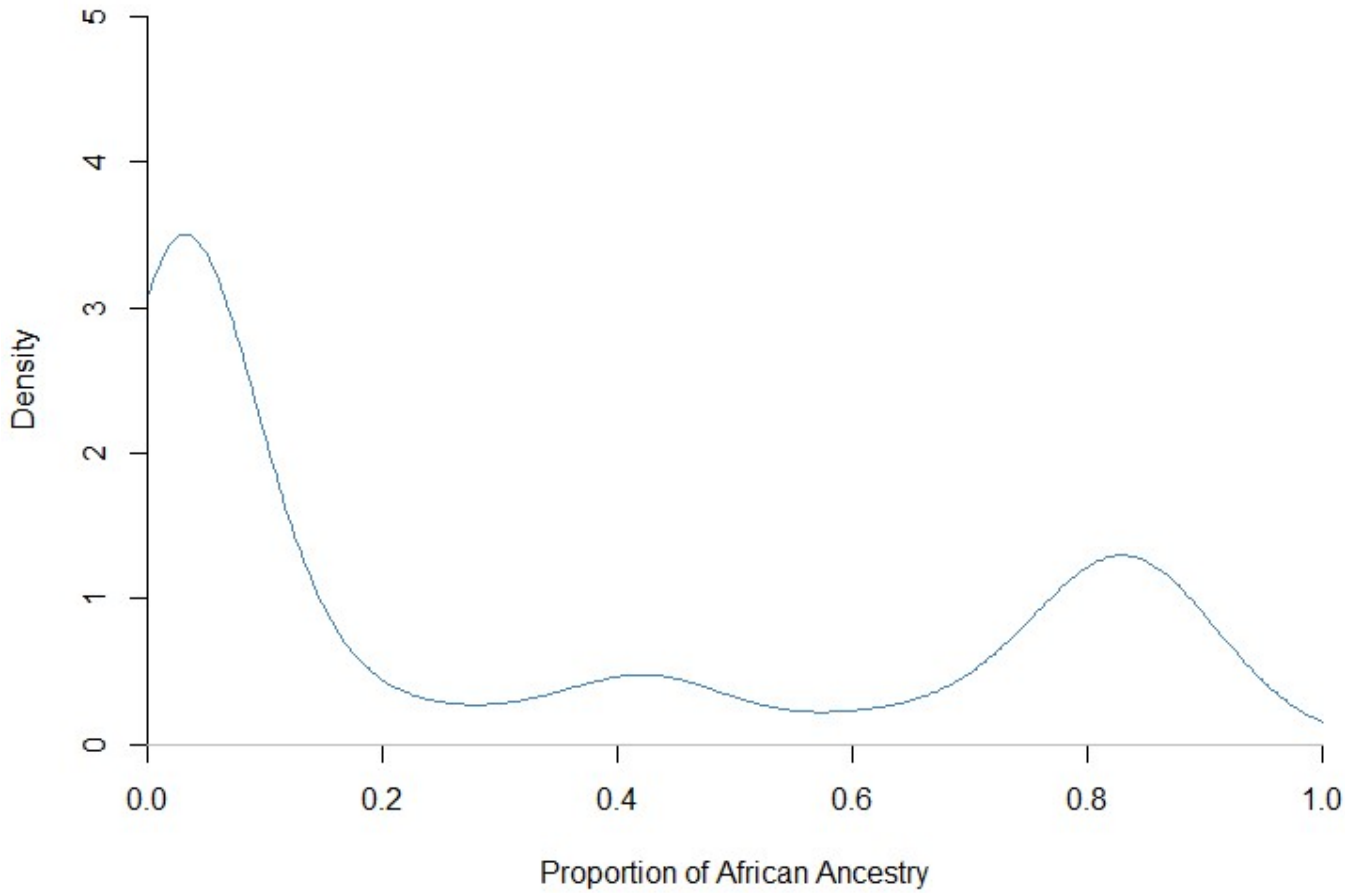

Figure 3

Linear and Nonlinear Gradients Measuring the Impact of African Ancestry

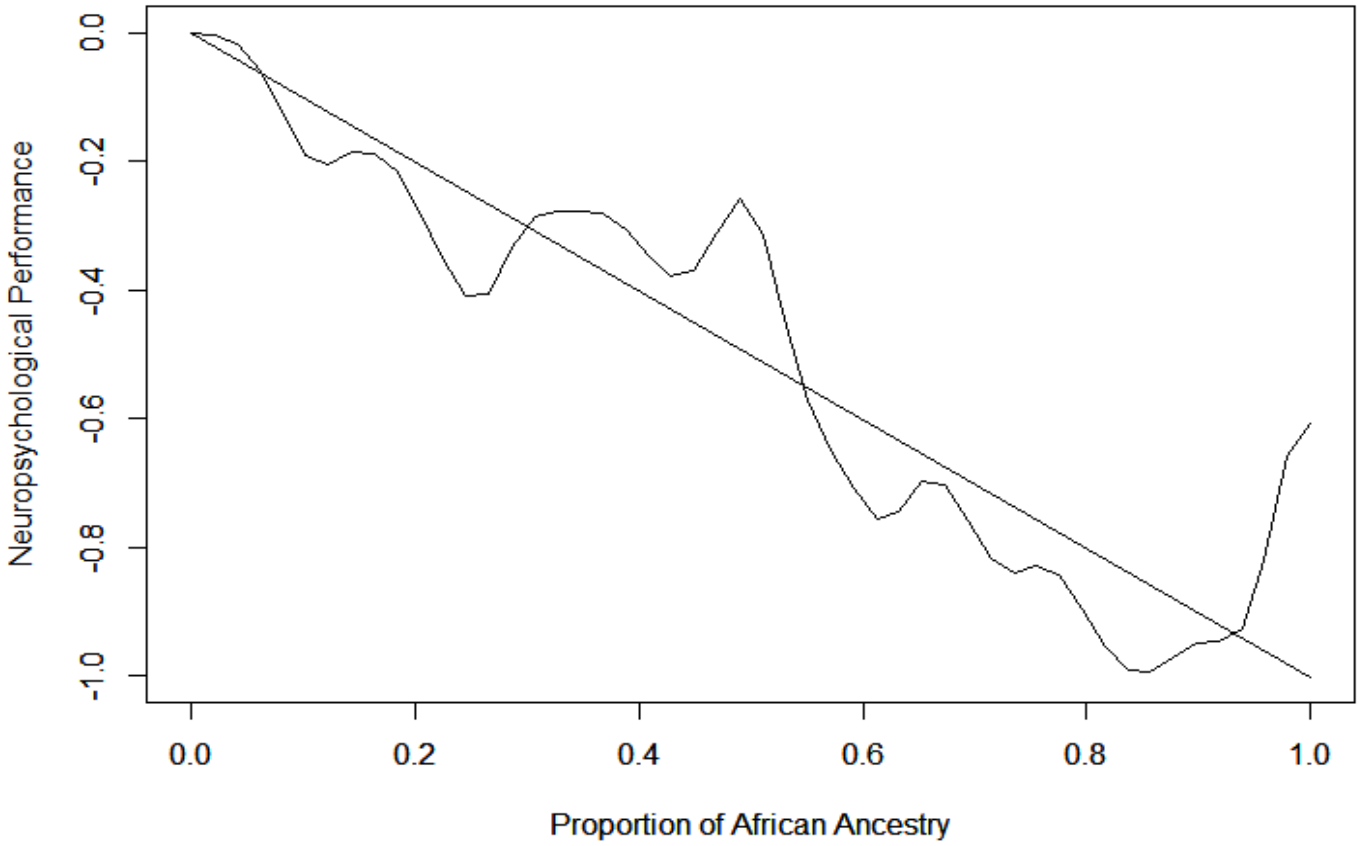

Table 4

Piecewise Linear Admixture Regression Results with and without Composite Groups

	Core Explanatory Variables										
	Intercept	% African	% Amerindian	% East Asian	% South Asian	Black SIRE	Hispanic SIRE	Native American SIRE	East Asian SIRE	South Asian SIRE	Other SIRE
Model 4	0.3017	-1.0791	-1.3897	0.6365	0.4776	-0.1132	-0.0819	-0.1351	-0.2049	-0.0668	-0.1882
t-statistic	9.1120	-8.5290	-11.4390	3.1480	1.4630	-1.1060	-0.9690	-1.2610	-1.1650	-0.2560	-2.4080
Model 5	0.2956	-1.0949	-1.3437	0.6501	0.5316	-0.1163	0.3501	-0.1712	-0.3418	-0.2584	-0.1650
t-statistic	8.9890	-8.4750	-10.8440	3.2030	1.6220	-1.1160	1.7810	-1.2650	-1.8700	-0.9580	-1.6170
	Piecewise Linear Variable	Multiple-SIRE-Composite Explanatory Variables									
	$D[A_2 \geq 0.9]A_2$										
Model 4 [cont.]	0.1598	Black-White SIRE	Hispanic-White SIRE	Native America - White SIRE	East Asian - White SIRE	South Asian - White SIRE	Hispanic - Black SIRE	Hispanic - Other SIRE			
t-statistic	1.8050										
Model 5 [cont.]	0.1809	0.0782	-0.0749	-0.0997	0.0085	0.3401	0.0700	-0.1566	Wald Test	Wald Test p-value	
t-statistic	2.0370	1.0300	-1.6450	-1.1820	0.0830	1.8220	0.5650	-2.2650			
Test 2		2.4208	-2.5439	-0.1355	2.4516	2.8161	-1.7625	-2.1549	26.3863	0.0004	
Conditional R2		Model 4: 0.550; Model 5: 0.549									
Marginal R2		Model 4: 0.157; Model 5: 0.160									

Notes to Table: Model 4 uses single-SIRE categories with multiple-SIRE choices allocated evenly across the categories; Model 5 adds seven multiple-SIRE categories. Both models include a kinked-linear explanatory variable for African ancestry above 90%. Test 2 gives the z-statistic for testing if the multiple-SIRE group coefficient equals the average of the component coefficients; the Wald statistic provides a joint test of all the Test 2 restrictions.

Table 5

Linear Admixture Regression Results Including Social-Economic Status (SES) and US Born Variables

Table 5a: Using Raw SES

	Core Explanatory Variables										
	Intercept	% African	% Amerindian	% East Asian	% South Asian	Black SIRE	Hispanic SIRE	Native American SIRE	East Asian SIRE	South Asian SIRE	Other SIRE
Model 6a	0.0766	-0.6111	-0.8174	0.6225	0.4315	-0.1027	-0.0309	-0.0360	-0.2144	-0.1124	-0.0722
t-statistic	1.2420	-4.9230	-6.8510	3.1730	1.3600	-1.0350	-0.3780	-0.3470	-1.2560	-0.4450	-0.9500
Model 7a	0.0810	-0.6335	-0.7791	0.6297	0.4706	-0.1031	0.3705	-0.0495	-0.3079	-0.2805	-0.0372
t-statistic	1.3150	-5.0010	-6.4070	3.1970	1.4780	-1.0210	1.9450	-0.3760	-1.7350	-1.0730	-0.3740
	Piecewise Linear Variable	Socio-Economic Variables		Multiple-SIRE-Composite Explanatory Variables							
	$D[A_2 \geq 0.9]A_2$	Socio-Economic Status	US Born								
Model 6a [cont.]	0.0179	0.2846	0.1013	Black-White SIRE	Hispanic-White SIRE	Native America -White SIRE	East Asian - White SIRE	South Asian - White SIRE	Hispanic - Black SIRE	Hispanic - Other SIRE	
t-statistic	0.2080	24.0880	1.7800								
Model 7a [cont.]	0.0422	0.2841	0.0909	0.1164	-0.0419	-0.0618	-0.0347	0.2836	0.0624	-0.0756	Wald test and p-value
t-statistic	0.4880	24.0590	1.5960	1.5830	-0.9520	-0.7570	-0.3500	1.5700	0.5200	-1.1260	25.6507
Test 2				3.2033	-1.7969	-0.8785	3.0760	-2.3854	-0.3681	1.6719	0.0006

Table 5b: Using Orthogonalized SES

	Core Explanatory Variables										
	Intercept	% African	% Amerindian	% East Asian	% South Asian	Black SIRE	Hispanic SIRE	Native American SIRE	East Asian SIRE	South Asian SIRE	Other SIRE
Model 6b	0.1980	-1.1207	-1.3573	0.7187	0.6246	-0.1192	-0.0863	-0.2418	-0.2122	-0.1179	-0.1739
t-statistic	3.2270	-9.1420	-11.5920	3.6630	1.9690	-1.2020	-1.0580	-2.3340	-1.2430	-0.4670	-2.2910
Model 7b	0.2026	-1.1358	-1.3115	0.7320	0.6759	-0.1236	0.3573	-0.2693	-0.3488	-0.3146	-0.1590
t-statistic	3.3060	-9.0710	-10.9760	3.7160	2.1220	-1.2230	1.8760	-2.0510	-1.9650	-1.2030	-1.6010
	Piecewise Linear Variable	Socio-Economic Variables		Multiple-SIRE-Composite Explanatory Variables							
	$D[A_2 \geq 0.9]A_2$	Socio-Economic Status	US Born								
Model 6b [cont.]	0.2081	0.2846	0.1013	Black-White SIRE	Hispanic-White SIRE	Native America -White SIRE	East Asian - White SIRE	South Asian - White SIRE	Hispanic - Black SIRE	Hispanic - Other SIRE	
t-statistic	2.4230	24.0880	1.7800								
Model 7b [cont.]	0.2295	0.2841	0.0909	0.0743	-0.0806	-0.1631	-0.0025	0.3185	0.0814	-0.1499	Wald test and p-value
t-statistic	2.6650	24.0590	1.5960	1.0110	-1.8340	-1.9970	-0.0250	1.7630	0.6770	-2.2360	28.5587
Test 2				2.7864	-2.5972	-0.8024	2.4965	-2.7226	-0.2826	2.4080	0.0002
Conditional R2	Models 6a,b: 0.548; Models 7a,b: 0.547										
Marginal R2	Models 6a,b: 0.219; Models 7a,b: 0.222										

Notes to Table: Models 6a,b use single-SIRE categories with multiple-SIRE choices allocated evenly among them; Models 7a,b add seven multiple-SIRE categories. All models include a kinked-linear explanatory variable for African ancestry above 90%, a dummy variable for a child born in the US, and a composite variable measuring social-economic status. In Models 6a and 7a the social-economic status variable is in raw form whereas in Models 6b and 7b it is orthogonalized with respect to the other explanatory variables (except US born). Test 2 gives the z-

statistic for testing if the multiple-SIRE group coefficient equals the average of the component coefficients; the Wald statistic provides a joint test of all the Test 2 restrictions.

Updated June 5

Supplementary material for “Linear and partially linear models of behavioral trait variation using admixture regression” in which the dataset, variables, and methods for the empirical analysis are detailed.

1. Materials and Methods

1.1 Dataset

The Adolescent Brain Cognitive Development Study (ABCD) is a collaborative longitudinal project between 21 sites across the US. Its goal is to further research into the psychological and neurobiological basis of development. At baseline, around 11,000 9-10 year old children were sampled, using a probabilistic sampling strategy, from public and private elementary schools and through non-school-based community outreach between 2016 and 2018, with the goal of creating a broadly representative sample of US children of this age. Children who were not fluent in English (or whose parents were not fluent in either English or Spanish) were excluded, along with those with severe medical, neurological, or psychiatric conditions. Informed consent was provided by parents.

Baseline ABCD 3.0 data release was used. For this analysis, we excluded individuals who did not have NIH Toolbox® results, who did not have admixture data, and who were identified as being a Pacific Islander. This left 9972 individuals.

1.2 Variables

1.2.1. Admixture

Subjects were genotyped using Illumina XX, with 516,598 variants directly genotyped and surviving the quality control done by the data provider. We used the 3.0 release of the dataset, which also includes an edition with imputed variants using TOPMED and Eagle 2.4. Because we had very few samples from Pacific Islanders, we excluded these from further analysis to simplify the reference

populations needed (n = 69). All our work was done on build 38. Files in hg17/37 were lifted to hg38 using liftOver (https://github.com/sritchie73/liftOverPlink) and the GRC chain file at ftp://ftp.ensembl.org/pub/assembly_mapping/homo_sapiens/ (GRCh37_to_GRCh38.chain.gz).

Before global admixture estimation, we applied quality control using plink 1.9. We used only directly genotyped, bi-allelic, autosomal SNP variants (494,433, 493,196, before and after lifting). We pruned variants for linkage disequilibrium at the 0.1 R^2 level using plink 1.9 (--indep-pairwise 10000 100 0.1), as recommended in the admixture documentation (https://vcru.wisc.edu/simonlab/bioinformatics/programs/admixture/admixture-manual.pdf). This variant filtering was done in the reference population dataset to reduce bias from sample representativeness. After pruning, we were left with 99,642 variants. To ensure a reasonable balance in the estimation dataset, we merged the target samples from ABCD, with reference population data for the populations of interest. We desired a k=5 solution (European, Amerindian, African, East Asian, and South Asian), so we merged with relevant samples from 1000 Genomes and from the HGDP. The following populations were excluded: Adygei, Balochi, Bedouin, Bougainville, Brahui, Burusho, Druze, Hazara, Makrani, Mozabite, Palestinian, Papuan, San, Sindhi, Uygur, Yakut. These reference populations were excluded because they were overly admixed or because, in the case of Melanesians and San, the individuals in the ABCD sample lacked significant portions of these ancestries.

Because the estimation sample would still be very skewed towards European ancestry using this joint sample, we used repeated subsetting to achieve balance. Specifically, we split the ABCD target samples into 50 random subsets, each with about 222 persons, and merged them one at a time with the reference data, followed by running admixture k=5 on each merged subset. We verified that these subsets produced stable results by examining the stability of the estimates for the reference samples. There was very little variation across runs, e.g. for the reference sample with the most

variance (European, NA12342), the mean estimate was 98.3% with SD=0.17% across the 50 runs. Since Admixture does not label the resulting clusters, we used 5 reference samples to index the populations so the data would be merged correctly. In no case did this produce any inconsistencies.

1.2.2. Neuropsychological Performance

The NIH Toolbox® (NIHTBX) neuropsychological battery was designed to measure a broad range of cognitive abilities. It consists of seven tasks which index attention (Flanker Inhibitory Control and Attention Task), episodic memory (Picture Sequence Memory Task), language abilities (Picture Vocabulary Task & Oral Reading Recognition Task), executive function (Dimensional Change Card Sort Task & Flanker Inhibitory Control and Attention Task), processing speed (Pattern Comparison Processing Speed Task), and working memory (List Sorting Working Memory Task) (Akshoomoff et al., 2014; Weintraub et al., 2013). NIHTBX was normed for samples between ages 3 and 85; tasks correlate highly with comparable ability assessmnents (Weintraub et al., 2013). Moreover, this battery has been shown to be measurement invariant across American ethnic groups (Lasker, Pesta, Fuerst, & Kirkegaard, 2019).

Age-corrected composite scores, based on the seven tasks, were provided by ABCD. We regressed out sex from these age-corrected composite scores. The residuals were then standardized.

1.2.3. Self-identified Race and Ethnicity

Self-identified race was based on parental responses to 18 questions asking about the child's race ("What race do you consider the child to be? Please check all that apply"). From these questions, six broad racial categories were created: European ("White"), African ("Black/African American"), Native American ("American/Native American" and "Alaska Native"), South Asian ("Asian Indian"), East Asian ("Chinese," "Filipino," "Japanese," "Korean," and "Vietnamese", "Other Asian,"), and Other ("Other race," "Refused to answer," "Don't know"). The Other Asian group (N = 66) was

classified as “East Asian” because the Asian ancestry component was predominantly East (44%;) not South (7%) Asian; the remaining ancestry was predominantly European (40%). The Pacific Islander groups (“Native Hawaiian,” “Guamanian,” “Samoan,” and “Other Pacific Islander”) were excluded as we did not have a corresponding admixture component. Self-identified ethnicity was based on parental responses to 1 question asking about Latin American ethnicity (“Do you consider yourself Hispanic/Latino/Latina?”). From this we created an additional ethnic category.

Descriptive statistics for the SIRE groups are shown in Table S1. Statistics are reported for single ethnic categories i.e., individuals reported as being only White, Black, East Asian, Native American, or Other, with no combinations (e.g., Hispanic & White), Hispanics, the seven top double combinations (i.e., Hispanic & White, Hispanic & Black, Hispanic & Other, non-Hispanic Black & White, non-Hispanic East Asian & White, non-Hispanic Native American & White, and non-Hispanic South Asian & White) and finally all other remaining groups combined.

Table 1. Descriptive Statistics for the SIRE Groups.

	N	Age *M*	Eur. %	Afr. %	E.Asian %	S.Asian %	Amer. %	US Born *N*	Neuro psy.	SES
Total Sample	9972	9.91 (0.63)	0.74 (0.31)	0.16 (0.29)	0.03 (0.12)	0.01 (0.06)	0.06 (0.13)	9703	0.00 (1.00)	0.00 (1.00)
NH White Only	5533	9.93 (0.63)	0.97 (0.05)	0.01 (0.02)	0.00 (0.02)	0.01 (0.02)	0.01 (0.03)	5459	0.25 (0.92)	0.40 (0.75)
NH Black Only	1434	9.91 (0.61)	0.18 (0.11)	0.80 (0.11)	0.00 (0.02)	0.00 (0.01)	0.01 (0.02)	1402	-0.77 (0.87)	-1.01 (0.94)
NH East Asian Only	107	10.02 (0.62)	0.14 (0.18)	0.01 (0.05)	0.82 (0.23)	0.02 (0.10)	0.01 (0.02)	88	0.57 (1.02)	0.58 (0.77)
NH South Asian Only	43	10.03 (0.68)	0.24 (0.13)	0.00 (0.00)	0.03 (0.07)	0.73 (0.14)	0.01 (0.01)	35	0.45 (1.02)	0.87 (0.45)

NH Native American Only	31	9.70 (0.60)	0.71 (0.30)	0.11 (0.26)	0.01 (0.02)	0.01 (0.01)	0.15 (0.19)	31	-0.42 (0.79)	-0.78 (0.73)
NH Other Only	97	9.96 (0.61)	0.55 (0.30)	0.28 (0.31)	0.06 (0.19)	0.04 (0.11)	0.07 (0.15)	87	-0.22 (1.12)	-0.56 (1.12)
Any Hispanic	1869	9.88 (0.63)	0.60 (0.20)	0.10 (0.14)	0.02 (0.06)	0.01 (0.02)	0.27 (0.18)	1755	-0.23 (0.96)	-0.41 (0.92)
White Hispanic	1171	9.89 (0.64)	0.67 (0.18)	0.06 (0.06)	0.01 (0.02)	0.01 (0.01)	0.26 (0.17)	1097	-0.19 (0.96)	-0.29 (0.91)
Black Hispanic	84	9.80 (0.64)	0.35 (0.13)	0.53 (0.17)	0.00 (0.01)	0.00 (0.01)	0.11 (0.09)	79	-0.34 (0.93)	-0.65 (1.00)
Other Hispanic	411	9.89 (0.63)	0.49 (0.15)	0.09 (0.11)	0.02 (0.03)	0.00 (0.01)	0.40 (0.17)	383	-0.45 (0.92)	-0.74 (0.87)
NH Black & White	302	9.88 (0.62)	0.58 (0.12)	0.41 (0.12)	0.00 (0.01)	0.00 (0.01)	0.01 (0.02)	301	-0.13 (0.95)	-0.45 (1.06)
NH East Asian & White	249	9.99 (0.64)	0.56 (0.12)	0.01 (0.01)	0.41 (0.14)	0.02 (0.04)	0.01 (0.01)	244	0.58 (0.98)	0.67 (0.67)
NH Native American & White	131	9.78 (0.60)	0.90 (0.10)	0.01 (0.03)	0.01 (0.02)	0.01 (0.01)	0.07 (0.09)	131	0.01 (0.86)	-0.07 (0.91)
NH South Asian & White	40	9.79 (0.54)	0.63 (0.11)	0.00 (0.00)	0.02 (0.08)	0.34 (0.12)	0.01 (0.01)	39	0.83 (0.84)	0.78 (0.78)
Any_Other NH combination	136	9.87 (0.62)	0.37 (0.21)	0.46 (0.25)	0.13 (0.20)	0.02 (0.07)	0.02 (0.05)	131	-0.27 (1.06)	-0.65 (1.03)

Note: Euro.% = European ancestry percentage, Afr.% = African ancestry percentage, E.Asian% = East Asian Ancestry percentage, S.Asian% = South Asian Ancestry percentage, Neuropsy. = Neuropsychiatric performance, SES = general socioeconomic component score. NH = non-Hispanic. Hispanic & White, Hispanic & Black, and Hispanic & Other are subsets of the Any Hispanic category.

The racial and ethnic variables were then recoded to create interval categories for which individuals are assigned a percentage of each SIRE category based on the number of responses chosen (Liebler & Halpern-Manners, 2008; Kirkegaard et al., 2019). By this coding, if someone was marked as White and Hispanic, they were assigned scores of .5 for white and .5 for Hispanic and 0 for the other 5 categories. The correlations between these interval scores and genetic ancestry components are shown in Table S2 below (*N* = 9972). These associations are similar to those found by others (for example: Guo et al. 2014); self-identified race generally corresponds with genetic ancestry.

Table S2. Correlations between Interval Coded SIRE and Genetic Ancestry.

	European ancestry	African ancestry	East Asian ancestry	South Asian ancestry	Amerindian ancestry
White_SIRE	**0.91**	-0.74	-0.21	-0.05	-0.31
Black_SIRE	-0.76	**0.96**	-0.08	-0.09	-0.18
East_Asian_SIRE	-0.24	-0.08	**0.92**	0.02	-0.06
South_Asian SIRE	-0.11	-0.04	0.01	**0.87**	-0.03
Native_SIRE	-0.02	0.00	-0.01	-0.03	0.07
Hispanic_SIRE	-0.21	-0.10	-0.04	-0.06	**0.77**
Other_SIRE	-0.16	-0.01	0.00	0.00	**0.38**

1.2.4. Region of Birth (US Born)

Region of birth was based on the parental response to the question, “In which country was the child born?”. The response “United States” was recoded as 1 and all other responses were recoded as 0.

1.2.5. Socioeconomic Status

Socioeconomic status was based on seven indicators: financial adversity, area deprivation index, neighborhood safety protocol, parental education, parental income, parental marital status, and parental employment status. These are detailed below:

1.2.5.1. Financial Adversity (Reverse Coded). Parents answered a seven item Financial Adversity Questionnaire (PRFQ). They were asked: “In the past 12 months, has there been a time when you and your immediate family experienced any of the following:

(1) “Needed food but could not afford to buy it or could not afford to go out to get it?”,
(2) “Were without telephone service because you could not afford it?”
(3) “Did not pay the full amount of the rent or mortgage because you could not afford it?”,
(4) “Were evicted from your home for not paying the rent or mortgage?”,
(5) “Had services turned off by the gas or electric company, or the oil company would not deliver oil because payments were not made?”,
(6) “Had someone who needed to see a doctor or go to the hospital but did not go because you could not afford it?”, and
(7) “Had someone who needed a dentist but could not go because you could not afford it?”
For each of the seven items they answered “yes” (1) or “no” (0). We summed responses. Thus the maximum was 7 and the minimum was 0.

This variable was reverse coded, so that higher scores indicated less financial adversity, and then standardized.

1.2.5.2. Area Deprivation Index (ADI) (Reverse Coded). Parents completed a residential history questionnaire. They provided the residential addresses and the number of full years they lived at each residence. For each address an Area Deprivation Index (ADI) was computed by ABCD and the

national percentile of the area's socioeconomic status was given. ADI was based on the following variables:

1. "Percentage of occupied housing units without complete plumbing (log)"
2. "Percentage of occupied housing units without a telephone"
3. "Percentage of occupied housing units without a motor vehicle"
4. "Percentage of single"
5. "Percentage of population below 138% of the poverty threshold"
6. "Percentage of families below the poverty level"
7. "Percentage of civilian labor force population aged >=16 y unemployed (unemployment rate)"
8. "Percentage of occupied housing units with >1 person per room (crowding)"
9. "Percentage of owner"
10. "Median monthly mortgage"
11. "Median gross rent"
12. "Median home value"
13. "Income disparity defined by Singh as the log of 100 x ratio of the number of households with <10000 annual income to the number of households with >50000 annual income"
14. "Median family income"
15. "Percentage of population aged >=25 y with at least a high school diploma"
16. "Percentage of population aged >=25 y with <9 y of education"

We weighted the ADI for the last three residences by the numbers of years at each residence. Before weighting, we recoded zero full years of residence as one-half of a year, so to give weight to time spend at a residence that was less than one year. The weighted ADI scores were then reverse coded, so that higher values indicated higher socioeconomic neighborhoods, and then standardized..

1.2.5.3. Neighborhood Safety Protocol. Parents were asked three Likert scale (1 = strongly disagree; 5 = strongly agree) questions about neighborhood safety: "I feel safe walking in my neighborhood, day or night," "Violence is not a problem in my neighborhood," and "My neighborhood is safe from crime." We used the precomputed summary scores for which the three scores were summed and then divided them by three.

1.2.5.3. Education. Parents were asked, "What is the highest grade or level of school you have completed or the highest degree you have received." To create an interval variable, we recoded

parental education as 0 to 18: Never attended/Kindergarten only = 0, 1st grade = 1, 2nd grade = 2, 3rd grade = 3, 4th grade = 4, 5th grade = 5, 6th grade = 6, 7th grade = 7, 8th grade = 8, 9th grade = 9, 10th grade = 10, 11th grade = 11, 12th grade = 12, High school graduate =12, GED or equivalent Diploma General =12, Associate degree: Occupational Program =14, Associate degree: Academic Program = 14, Bachelor's degree = 16, Master's degree = 18, Professional school = 18, Doctoral degree = 18. We standardized the scores for each educational scores for both parents and standardized the averaged scores.

1.2.5.4. Income. Family was an interval variable which reflected the parents' total combined family income in the past 12 months. The variable was recoded as follows: 1.00 = less than $5000 (recode: 4,500); 2.00 = $5000 to 11,999 (recode: 5,000); 3.00 = $12,000 to 15,999 (recode: 12,000); 4.00 = $16,000 to 24,999 (recode: 16,000); 5.00 = $25,000 to 34,999 (recode: 25,000); 6.00 = $35,000 to 49,999 (recode: 35,000); 7.00 = $50,000 to 74,999 (recode: 50,000); 8.00 = $75,000, to 99,999 (recode: 75,000); 9.00 = $100,000 to 199,999 (recode: 100,000); 10.00 = $200,000 and greater (recode: 200,000).

1.2.5.5. Marital Status. Parental marital status was coded as 1 if married and 0 for any other arrangement.

1.2.5.6. Employment Status. Parental employment was coded as 1 if at least one parent was working now either full or part time and 0 for all other cases.

3.2.5.7. General SES. Missing data for the seven economic indicators were imputed using the mice package (df, m=5, maxit = 50, method = 'pmm', seed = 500). Descriptive statistics for the

imputed SES indicators are provided in Table S3, while the correlation matrix for the imputed variables (N = 9972), along with neuropsychiatric performance, is shown in Table S4.

Table S3. Descriptive Statistics for the SES Indicators.

Variable	Mean	SD	Median	Min	Max	Skew	Kurtosis	S.E.
Financial Adversity	0.00	1.00	0.42	-6.00	0.42	-2.90	8.90	0.01
ADI	0.00	1.00	0.22	-2.26	1.42	-0.64	-0.62	0.01
Safety Protocol	0.00	1.00	0.09	-3.04	1.13	-0.91	0.44	0.01
Education	0.00	1.00	-0.02	-5.44	1.33	-0.82	1.19	0.01
Income	0.00	1.00	-0.05	-1.33	2.24	0.92	0.35	0.01
Marital Status	0.68	0.47	1.00	0.00	1.00	-0.78	-1.39	0.00
Employment Status	0.91	0.28	1.00	0.00	1.00	-2.93	6.58	0.00

Table S4. Correlation Matrix for the SES indicators, Genetic Ancestry, and Neuropsychiatric Performance (N = 9972 in all cases).

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
(1) Financial Adversity	1.00												
(2)ADI	0.23	1.00											
(3)Safety Protocol	0.24	0.26	1.00										
(4)Education	0.29	0.30	0.28	1.00									
(5)Income	0.34	0.36	0.32	0.57	1.00								
(6)Marital Status	0.27	0.25	0.23	0.33	0.46	1.00							
(7)Employed	0.16	0.14	0.15	0.26	0.27	0.28	1.00						
(8)European	0.24	0.29	0.30	0.38	0.38	0.37	0.21	1.00					
(9)African	-0.26	-0.35	-0.28	-0.29	-0.36	-0.40	-0.20	-0.82	1.00				
(10)Amerindian	-0.06	0.00	-0.12	-0.39	-0.23	-0.09	-0.08	-0.28	-0.13	1.00			
(11)East_Asian	0.05	0.07	0.02	0.10	0.09	0.07	0.02	-0.28	-0.09	-0.04	1.00		
(12)South_Asian	0.04	0.05	0.04	0.12	0.11	0.07	0.03	-0.07	-0.10	-0.07	0.00	1.00	
(13)Neuropsychic Performance	0.20	0.23	0.18	0.40	0.35	0.27	0.18	0.33	-0.35	-0.15	0.11	0.09	1.00

We then submitted the seven SES indicators to Principal Component Analysis (PCA). For this, we used the R package PCAmixdata, which handles mixed categorical and continuous data (Chavent, Kuentz-Simonet, & Saracco, 2014). The first unrotated component explained 40% of the variance. The PCA_1 loadings for the seven SES indicators were as follows: financial adversity (.250), area deprivation index (.153), neighborhood safety protocol (.256), parental education (.504), parental income (.658), parental marital status (.270), and parental employment status (0.183). The vector of PCA_1 loadings correlated with the vector of SES indicator effects on Neuropsychiatric Performance at .84 (*N*=7). This indicates that the better measures of general SES have higher cognitive loadings.

PCA_1 scores correlated with Neuropsychological Performance at r = .43 in the full sample. Among non-Hispanic Whites, non-Hispanic Blacks, and Hispanics, the correlations between PCA_1 and Neuropsychological Performance was r = .26 (N = 5533), r = .34 (N = 1434), and r = .32 (N = 1869), respectively. These magnitudes of child-parental SES correlations are consistent with those previously reported (Flores-Mendoza, Ardila, Gallegos, & Reategui-Colareta, 2021; Sirin, 2005). The congruence coefficients for the SES component loadings were greater or equal to r = .97 for the largest three SIRE groups (non-Hispanic Whites, non-Hispanic Blacks, and Hispanics), indicating identical structures across groups.

2. Methods

A series of regression models were run with NIHTBX as the dependent variable. The NIHTBX and socioeconomic variables were standardized (based on the subsample of 9972 retained). The ancestry variables were left unstandardized, thus the coefficients from ancestries can be interpreted as a change in 100% ancestry over a change in one standardized unit of NIHTBX scores. European ancestry and White SIRE were selected as reference values and thus not included as independents.

For the regression analyses, following the recommendations of Heeringa and Berglund (2021), we used a three-level (site, family, individual) multi-level mixed effects model. This model was applied to the pooled twin and regular ABCD baseline sample. This specification approximates the ABCD Data Exploration and Analysis Portal (DEAP) specification (Heeringa and Berglund, 2021).

References

Akshoomoff, N., Newman, E., Thompson, W. K., McCabe, C., Bloss, C. S., Chang, L., ... & Jernigan, T. L. (2014). The NIH Toolbox Cognition Battery: results from a large normative developmental sample (PING). *Neuropsychology*, 28(1), 1.

Chavent, M., Kuentz-Simonet, V., & Saracco, J. (2014). Multivariate analysis of mixed data: the PCAmixdata R package. arXiv. arXiv preprint arXiv:1411.4911.

Flores-Mendoza, C., Ardila, R., Gallegos, M., & Reategui-Colareta, N. (2021) General Intelligence and Socioeconomic Status as Strong Predictors of Student Performance in Latin American Schools: Evidence From PISA Items. *Front. Educ*. 6:632289. doi: 10.3389/feduc.2021.632289

Guo, G., Fu, Y., Lee, H., Cai, T., Harris, K. M., & Li, Y. (2014). Genetic bio-ancestry and social construction of racial classification in social surveys in the contemporary United States. *Demography*, 51(1), 141-172.

Heeringa, S. G., & Berglund, P. A. (2020). A guide for population-based analysis of the Adolescent Brain Cognitive Development (ABCD) Study baseline data. *BioRxiv*.

Kirkegaard, E. O., Woodley Of Menie, M. A., Williams, R. L., Fuerst, J., & Meisenberg, G. (2019). Biogeographic ancestry, cognitive ability and socioeconomic outcomes. *Psych*, 1(1), 1-25.

Lasker, J., Pesta, B. J., Fuerst, J. G., & Kirkegaard, E. O. (2019). Global ancestry and cognitive ability. *Psych*, 1(1), 431-459.

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of educational research*, 75(3), 417-453.

Weintraub, S., Dikmen, S. S., Heaton, R. K., Tulsky, D. S., Zelazo, P. D., Bauer, P. J., ... & Gershon, R. C. (2013). Cognition assessment using the NIH Toolbox. *Neurology*, 80(11 Supplement 3), S54-S64.