

Article

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

Gregory Connor ^{1*}  and Gerard R. Fuerst ²

¹ Maynooth University

² Cleveland State University

* Correspondence: gregory.connor@mu.ie

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 racial and ethnic groups. This paper provides a statistical framework for admixture regression based on the linear polygenic index model and applies it to neuropsychological performance data from the Adolescent Brain Cognitive Development (ABCD) database. We develop and apply 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. We find a statistically significant genetic component to neuropsychological performance differences across racial identities, and find some possible evidence of nonlinearity in the link between admixture and neuropsychological performance scores in the ABCD data.

Keywords: mixed effects model; orthogonalized regressors; partially linear semiparametric regression; polygenic index

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 behavioral 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 behaviors (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 behavioral traits including asthma (Salari et al. 2005), 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 behavioral 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 conjunction with admixture regression, which uses ancestral

Citation: Connor, G.; Fuerst, G. Admixture Regression. *Behav. Sci.* **2021**, *1*, 0. <https://doi.org/>

Received:

Accepted:

Published:

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Submitted to *Behav. Sci.* for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

38 population trait differences to attempt to identify genetic loci associated with a trait.
39 This paper focusses exclusively on admixture regression.

40 This paper first develops a simple statistical framework for admixture regression
41 of behavioral traits by linking it to the linear polygenic index model from behavioral
42 genetics; this framework clarifies the key assumptions that are implicit in this simple and
43 powerful statistical technique. The paper then extends the admixture regression method-
44 ology in several ways. We provide a new test statistic for identifying whether a given
45 multi-racial identity differs in its trait impact from the average impact of its component
46 single-SIRE categories. We examine the role of additional explanatory variables in the
47 admixture regression and their interpretation with and without orthogonalization with
48 respect to the core explanatory variables. We generalize the linear admixture regression
49 specification to a partially linear semiparametric form.

50 We apply our methodology to neuropsychological performance data from the
51 Adolescent Brain Cognitive Development database. Neuropsychological performance is
52 one of the most complex traits to which admixture regression analysis has been applied.
53 Our findings corroborate existing evidence that genetic variation plays a statistically
54 significant role in explaining neuropsychological performance differences across racial
55 identities (Lasker et al. 2019). Using our new test statistic, we find that some multi-
56 racial categories have identifiably distinct impact relative to their component categories.
57 We find that orthogonalization of additional variables can substantially change the
58 interpretation of the core coefficients in the admixture regression. Our analysis also
59 indicates (although not conclusively) that a partially linear semiparametric specification
60 potentially adds empirical value.

61 2. A statistical framework for admixture regression tests of trait variation

62 2.1. Variable definitions

63 We assume that the database consists of n individuals indexed by $i = 1, \dots, n$ who
64 have each self-identified their racial or ethnic group membership(s), recorded a score
65 on a behavioral trait, s_i , and provided a personal DNA sample. The k racial or ethnic
66 group self-identification choices are captured by a matrix of zero-one dummy variables
67 $SIRE_{ij}$, $i = 1, \dots, n$; $j = 1, \dots, k$. We assume that every individual has self-identified as
68 belonging to at least one and possibly more of the k groups.

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

76 In most applications of admixture regression, individuals' racial or ethnic group
77 identities will have statistical relationships with individuals' genetically identified ge-
78 ographic ancestries and also with the observed trait s_i . The objective of admixture
79 regression is to decompose trait variation into linear components due to genetic ances-
80 tries and linear components due to racial/ethnic group related effects.

81 2.2. Ancestry proportions as a statistical proxy for ancestry-linked genetic trait variation

82 Admixture regression is an indirect method of analyzing group-related trait varia-
83 tion. In this subsection we provide a foundation for admixture regression by considering
84 a more direct, but empirically much more challenging, alternative approach based on
85 a linear polygenic index model. We show that the admixture regression model can be
86 viewed as a statistically feasible simplification of this linear polygenic index model, in
87 which proportional ancestries serve as statistical proxies for ancestry-related genetic
88 differences.

89 The human genetic code contains a very large number of genetic variants (the alleles
90 on the genome which vary between individuals) called single nucleotide polymorphisms
91 or SNPs. Consider hypothetically a complete list of all genetic variants with any impact
92 on variation in the observed trait. Assign a value of 0, 1 or 2 to each SNP for individual i
93 depending upon the number of minor alleles for that SNP. Let SNP_{iz} $i = 1, \dots, n$; denote
94 the number of minor alleles on the z^{th} SNP of the i^{th} individual in the sample.

95 The biochemical process linking human genetic variation to behavioral trait vari-
96 ation is unimaginably complex, and scientific understanding of the full biochemical
97 process is very limited. Genome-wide association studies (GWAS) have made slow but
98 steady progress in statistically modeling these linkages, although precise biochemical
99 linkages are beyond the contemporary scientific frontier for most behavioral traits. A
100 standard, admittedly highly simplified, model of the gene variation - trait variation
101 nexus is the linear polygenic index model, in which the genetic component of a trait is a
102 simple linear function of a relevant subset of the individual's genetic variants. The linear
103 polygenic index model has been applied to a wide range of medical and behavioral
104 traits including body mass index (Yengo et al. 2018), neuroticism (Nagel et al. 2018),
105 depression susceptibility (Wray et al. 2018), suicidal ideation (Mullins et al. 2014),
106 schizophrenia (Mistry et al. 2018), educational attainment (Lee et al. 2018), neuropsychol-
107 ological test performance (Savage et al. 2018), and risk-taking (Clifton et al. 2018). The
108 linear admixture regression model can be derived elegantly by invoking this standard
109 linear polygenic index model, and hence we impose it, in order to provide a statistical
110 underpinning for our admixture regression model.

Let p_i denote the genetic potential of individual i regarding the observable trait s_i .
We assume that p_i is a linear function of a large number of genetic variants SNP_{iz} with
associated linear coefficients β_z and constant term c_1 :

$$p_i = c_1 + \sum_z \beta_z \text{SNP}_{iz}; i = 1, \dots, n. \quad (1)$$

111 The key difference in the admixture regression methodology compared to GWAS is
112 that there is no attempt to estimate the linear polygenic index (1). Rather, admixture
113 regression uses the natural experiment of subpopulation mixing to infer differences in the
114 conditional expected value of (1) arising from differences in the frequency distribution of
115 genetic variants across ancestries. The assumed linearity of the polygenic index model
116 (1), together with an assumption of random mating across ancestral populations, allows
117 us to derive a linear regression model using admixture as a statistical proxy variable for
118 the conditional expected value of p_i .

The frequency distributions of many SNPs depend notably upon geographic an-
cestries. Consider a hypothetical individual with single-origin ancestry h , that is, an
individual with $A_h = 1$. Note that this also implies that $A_{h'} = 0$ for all $h' \neq h$ since the
ancestral proportions are non-negative and sum to one. Consider the expected value
of p conditional on an individual having this single-origin ancestry. The expectation of
 $\sum_z \beta_z \text{SNP}_z$ using a single-origin frequency distribution for each SNP_z defines the average
genetic trait potential of a single-origin ancestry:

$$E[p|A_h = 1] = c_1 + \sum_z \beta_z E[\text{SNP}_z|A_h = 1], h = 1, \dots, m. \quad (2)$$

119 In admixture regression there is no attempt to measure (2) directly, but instead differences
120 between (2) across $h = 1, \dots, m$ will be inferred indirectly using regression methods.

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 expected value of any SNP
for an admixed individual is the convex combination of the single-origin expected

values, with linear coefficients equal to the individual's admixture proportions. (The relationship between the multivariate distributions of the SNPs is more complicated, but the multivariate distributions do not impact the expected trait given the linear polygenic index assumption.) We use a subscript \cdot to denote the vector created from the i^{th} 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 single-origin frequency distributions:

$$E[\text{SNP}_{iz}|A_{i\cdot}] = \sum_{h=1}^m A_{ih}E[\text{SNP}_z|A_h = 1] \quad (3)$$

The linearity of genetic potential in the SNPs (1) and the random mixing assumption (3) imply that expected genetic potential of an admixed individual is a convex combination of the individual's admixture proportions. Taking the expectation of (1) using (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 $E[p|A_h = 1]$ with observed linear coefficients A_{ih} :

$$E[p_i|A_{i\cdot}] = \sum_{h=1}^m A_{ih}E[p|A_h = 1]. \quad (4)$$

121 Equation (4) is a fundamental identification condition for the admixture regression
 122 methodology. As we discuss below, it allows differences between the single-origin
 123 expected values of genetic potential, $E[p|A_h = 1]$, $h = 1, \dots, m$, to be inferred by regression
 124 methods.

125 2.3. Adjusting for ancestry-related environmental influences on the trait

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

$$s_i = p_i + e_i, \quad (5)$$

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 \tilde{p}_i as the genetic component of the trait for each i which is not explained by ancestry proportions:

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

by simple substitution into (5) this gives:

$$s_i = c_1 + \sum_{h=1}^m A_{ih}E[p|A_h = 1] + \tilde{p}_i + e_i. \quad (6)$$

Recall that $\sum_{h=1}^k A_{hi} = 1$, for all i , so that one term in (6) is redundant for the purposes of

creating a regression model. Substitute $A_{i1} = 1 - \sum_{h=2}^k A_{ih}$ into (6) to get:

$$s_i = c_2 + \sum_{h=2}^m b_{Ah}A_{ih} + \tilde{p}_i + e_i, \quad (7)$$

126 where $b_{Ah} = E[p|A_h = 1] - E[p|A_1 = 1]$; $h = 2, \dots, m$, and $c_2 = c_1 + E[p|A_1 = 1]$.

127 Equation (7) is not well-specified as a regression model since the error term $\tilde{p}_i + e_i$
 128 will not be mean zero conditional on $A_{i\cdot}$, due to racial and ethnic group-related effects in
 129 e_i . In order to transform (7) into a regression model it is necessary to add explanatory
 130 terms to the regression model to remove the expected value of e_i conditional on $A_{i\cdot}$. This

131 is accomplished by assuming that the differences in e_i conditional on A_i are dependent
 132 on the group self-identification choices, but otherwise not dependent upon admixture
 133 proportions.

134 For expositional simplicity, in this subsection we assume that every individual
 135 included in the sample has self-identified as belonging to exactly one from the pre-
 136 specified set of k racial or ethnic groups, so that $\sum_{j=1}^k \text{SIRE}_{ij} = 1$ for all i . In this case,
 137 the $n \times k$ matrix of racial/ethnic group explanatory variables used in the admixture
 138 regression, denoted G , is simply set equal to the SIRE matrix: $G_{ij} = \text{SIRE}_{ij}$ for $i =$
 139 $1, \dots, n; j = 1, \dots, k$. Multi-racial individuals (those who have self-identified as belonging
 140 to two or more groups) will be introduced into the analysis in the next subsection.

We assume that after adjusting for the influence of the group identifiers G_{ij} , the remaining error term in (7) is independent of the ancestry proportions:

$$e_i = c_3 + \sum_{j=2}^k b_{G_j} G_{ij} + \tilde{e}_i, \quad (8)$$

where b_{G_h} captures the environmental component associated with membership in group h relative to the reference group $h = 1$, and \tilde{e}_i is assumed to be independent of A_i , G_i , and \tilde{p}_i , and c_3 is a constant term. Combining (7) and (8) produces the key linear admixture regression specification:

$$s_i = c_4 + \sum_{j=2}^k b_{G_j} G_{ij} + \sum_{h=2}^m b_{A_h} A_{ih} + \varepsilon_i. \quad (9)$$

141 where $\varepsilon_i = \tilde{e}_i + \tilde{p}_i$ and c_4 is a constant term. Note that ε_i has zero mean and variance $\sigma_{\tilde{e}}^2 +$
 142 $\sigma_{\tilde{p}}^2$ and is independent of A_i and G_i . Equation (9) is a well-specified linear regression
 143 model.

144 In many applications, the analyst also has information on the sampling substructure
 145 of the data, such as its division into site-specific subsamples. In this case, a linear mixed
 146 effects model can be used for estimating (9) rather than ordinary least squares. This
 147 involves partially decomposing the residual term ε_i in (9) into linear random effects
 148 components linked to data collection site identifiers and/or other subsample identifiers,
 149 see Heeringa and Berglund (2020).

150 2.4. Adding multi-racial individuals to the regression

151 Recall that SIRE is the $n \times k$ matrix of race/ethnicity self-identifications. A key
 152 assumption of the admixture regression technique is that the environmental influences
 153 associated with racial/ethnic group membership are captured by these group member-
 154 ship self-identification choices. Many individuals self-identify as belonging to two or
 155 more racial or ethnic groups and the group variables used in the regression must be
 156 adapted to this reality. In the context of our statistical framework, there are essentially
 157 three approaches: evenly splitting the individual's affiliation across their chosen groups,
 158 creating a new group for one or more particular multi-racial combinations, or deleting
 159 particular multi-racial observations where neither of the other two approaches seem
 160 appropriate.

We now allow that some individuals choose more than one category, so that
 $\sum_{j=1}^k \text{SIRE}_{ij} > 1$ for some i . The simplest regression specification in this case is to as-

sume that the group environment faced by a multi-racial individual is the average of the component group environments:

$$G_{ij} = \text{SIRE}_{ij} / \left(\sum_{j^*=1}^k \text{SIRE}_{ij^*} \right) \text{ for all } i = 1, \dots, n; j = 1, \dots, k. \quad (10)$$

161

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

$$\begin{aligned} G_{ij} &= 0 \text{ for } j = 1, \dots, k \text{ if } \text{SIRE}_i = V_1 \\ G_{i,k+1} &= 1 \text{ if } \text{SIRE}_i = V_1 \\ G_{i,k+1} &= 0 \text{ if } \text{SIRE}_i \neq V_1, \end{aligned} \quad (11)$$

169 where $\text{SIRE}_i = V_1$ denotes vector equality between these two k -vectors. There are now
170 $k + 1$ groups: the originally specified SIRE groups and a new group for the selected
171 multiracial combination. G becomes a $n \times (k + 1)$ matrix, and the regression (9) described
172 in the previous subsection applies exactly as before but with one extra dimension to
173 G . Any small number of defined multi-racial groups can be appended in this way. The
174 only change to the regression methodology is that G becomes a $n \times k^*$ matrix (with an
175 associated increase in the set of estimated parameters) where $k^* - k$ is the number of
176 multiracial combinations added as new categories.

177 It is not feasible to use rule (11) for all race/ethnicity choice combinations due to
178 lack of parsimony; there are $2^k - k$ potential multi-racial combinations and each one
179 added requires an additional parameter in the regression. It can only be used for the
180 common multi-racial choices where there is sufficient data of that combination in the
181 sample. For all others, it is necessary to stick with the restrictive assumption (10) or drop
182 the observations from the sample. This will be illustrated in the empirical application in
183 Section 5.

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:

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

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

Let \hat{b} denote the $(m + k^* - 1)$ -vector of all the coefficients in the admixture regression (9):

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

189 and let $\widehat{\text{Cov}}_{\hat{b}}$ denote the estimated $(m + k^* - 1) \times (m + k^* - 1)$ -covariance matrix of
190 these estimates.

First consider the case $k^* - k = 1$. Let R denote the $(m + k^* - 1)$ -vector expressing restriction (12) imposed on b . For example, if the group combination consists of individ-

uals who choose all three of the first, second, and third SIRE categories (recalling that the first SIRE category is not included in the regression) the restriction vector is:

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

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:

$$\frac{\widehat{b}'R}{(R'\widehat{Cov}_{\widehat{b}}R)} \sim t(n - m - k^* + 1). \quad (13)$$

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 $(m + k^* - 1) \times (k^* - k)$ -matrix of all the linear restrictions, giving the standard Wald test:

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

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

194 3. Extensions of the linear admixture regression model

195 3.1. Additional explanatory variables with and without orthogonalization

It is straightforward to include additional explanatory variables in the admixture regression model. Let $x_{i1}, x_{i2}, \dots, x_{il}$ 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:

$$s_i = c_5 + \sum_{j=2}^k b_{Gj}G_{ij} + \sum_{h=2}^m b_{Ah}A_{ih} + \sum_{d=1}^l b_{xd}x_{id} + \varepsilon_i \quad (15)$$

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

198 In some cases, the admixture regression model with additional explanatory vari-
 199 ables (15) can be made more useful and informative by orthogonal rotation of one or
 200 more of the explanatory variables, in order to aggregate the full effects of proportional
 201 ancestries and group identities into their associated coefficients. To understand why such
 202 an orthogonal rotation might be useful, consider the hypothetical case of an admixture
 203 regression model of Body Mass Index (BMI) in which waist measurement is one of the
 204 explanatory variables. Waist measurement has such strong explanatory power for BMI
 205 that its presence in an admixture regression model like (15) will diminish the direct
 206 explanatory power of proportional ancestries and group identities; their total impact
 207 will be partly hidden within the waist measurement variable. This can be remedied
 208 by orthogonalizing the waist measurement variable with respect to the proportional
 209 ancestry and group identity variables before estimating the admixture regression, as
 210 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 variables, and therefore the analyst wishes to orthogonalize it with respect to G_{ij} and A_{ih} , $j = 2, \dots, k; h = 2, \dots, m$. 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 :

$$x_1 = \hat{c}_6 + \sum_{j=2}^k \hat{b}_{G_j} G_{ij} + \sum_{h=2}^m \hat{b}_{A_h} A_{ih} + x_1^o \quad (16)$$

211 Since all the explanatory variables are deterministic (that is, conditionally fixed variables
 212 rather than random variables in the regression model), this orthogonalization step is
 213 interpreted as a matrix transformation of fixed vectors and does not alter any statistical
 214 assumptions of the main regression model. It merely serves to linearly rotate the deter-
 215 ministic explanatory variables used in the actual, second-stage, admixture regression.
 216 Replacing x_1 with x_1^o in (15) changes the interpretation of the coefficients \hat{b}_{G_j} and \hat{b}_{A_h} ,
 217 $j = 2, \dots, k; h = 2, \dots, m$ since they now include the G_{ij} and A_{ih} related explanatory
 218 power from x_1 . An illustrative example will be provided in Section 5 below.

219 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:

$$p_i = p(\text{SNP}_{i.})$$

220 and instead of linearity as in (1) only require smoothness conditions on the conditional
 221 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 $A_{i.}$, but now without imposing the strict linearity (4) arising from the linear polygenic index assumption:

$$f(A_{i.}) = \mathbb{E}[p(\text{SNP}_{i.}) | A_{i.}].$$

Define the unexplained component of p_i as before:

$$\tilde{p}_i = p_i - f(A_{i.})$$

and we assume that $\tilde{p}_i \sim N(0, \sigma_{\tilde{p}}^2)$ and independent of $A_{i.}$ and $G_{i.}$. We impose the same assumptions on ε_i as in Section 2, giving:

$$s_i = \sum_{j=2}^k b_{G_j} G_{ij} + f(A_{i.}) + \varepsilon_i, \quad (17)$$

222 where $\varepsilon_i = \tilde{p}_i + \tilde{\varepsilon}_i$ is assumed to be normally distributed with mean zero and variance
 223 $\sigma_{\tilde{p}}^2 + \sigma_{\tilde{\varepsilon}}^2$ and independent of $A_{i.}$ and $G_{i.}$. This equation (17) is a partially linear nonpara-
 224 metric regression model, see, e.g. Li and Racine (2007). This model can be consistently
 225 estimated using the three-step procedure of Robinson (1988). We will impose Condition
 226 7.1 from Li and Racine (2007) in order to justify this procedure within our framework
 227 (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:

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

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

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

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

230 where $k(\bullet)$ is a kernel weighting function. In our empirical application in Section 5 we
231 use the Gaussian kernel weighting function.

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} :

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

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

and

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

that is:

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

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

and

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

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:

$$(\hat{b}_G, \hat{b}_A) = (X'X)^{-1} X'y$$

where

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

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

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

232 Note that (\hat{b}_G, \hat{b}_A) is a $(k + m - 3)$ -vector and X is a $n \times (k + m - 3)$ -matrix where the
233 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:

$$y_i^* = s_i - \left(\sum_{j=2}^k \hat{b}_{G_j} G_{ij} + \sum_{h=3}^m \hat{b}_{A_h} A_{ih} \right); i = 1, \dots, n$$

and then:

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

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

235 The partially linear nonparametric approach to admixture regression is more empiri-
236 cally challenging than the linear specification. Proper implementation of the technique
237 involves a tradeoff between parsimony, the generality of the specification used, and the
238 distributional features of the available data. An example of (18) will be estimated in
239 Section 5 below.

240 4. Materials and Methods

241 The Adolescent Brain and Cognitive Development (ABCD) study is the largest
242 long-term study of brain development and child health in the United States, testing
243 11,000 children ages 9-10 at 21 testing sites; see Karcher and Barch (2021) for an overview.
244 Our sample consists of age and gender-adjusted scores and genotyped DNA samples of
245 the 9972 children in the ABCD study who met our sample selection criteria, along with
246 questionnaire responses of their parent(s)/guardian(s).

247 The dependent variable in our model is the composite neuropsychological perfor-
248 mance score based on the NIH Toolbox (NIHTBX) neurocognitive battery provided in the
249 ABCD database; this consists of tasks measuring attention, episodic memory, language
250 abilities, executive function, processing speed, and working memory. Age-corrected
251 composite scores, based on the seven tasks, were provided by ABCD. We regressed out
252 sex from these age-corrected composite scores. The residuals were then standardized
253 and serve as the dependent variable in our empirical analysis in Section 5 below.

254 Our core explanatory variables are seven SIRE variables, White, Black, Hispanic,
255 Native American, East Asian, South Asian, and Other (and including multiple SIRE
256 choices from among these) and five genetic ancestry proportions of European, African,
257 Amerindian, East Asian and South Asian background obtained from the genotyped
258 DNA samples. Children whose parent(s)/guardian(s) identified the child as belonging
259 to Pacific Islander racial groups were excluded from our analyses owing to a lack of
260 corresponding ancestry category in our chosen five categories. The ABCD Version 3
261 database provides 516,598 genotyped SNP variants for each individual's DNA sample.
262 After quality control, filtering, and pruning we were left with 99,642 SNP variants to
263 determine the five ancestry proportions, employing the Admixture 1.3 software package
264 (Alexander et al. 2015). We use the Pritchard et al. (2000) population structure algorithm,
265 as implemented in R routine *Structure*, to estimate the ancestry proportions of each
266 individual in the sample.

267 The ABCD database includes site identifiers for the data collection sites (in most
268 cases, elementary schools) and family household identifiers (identifying multiple indi-
269 viduals in the sample from the same family household, usually twins). As recommended
270 by Heeringa and Berglund (2020) for regression analysis using the ABCD database, we
271 include random effects in our regression models to account for any site-specific and
272 family-specific error correlation. We use the linear regression mixed effects estimate
273 routine *lmer* from the R programming language library, see Bates et al. (2015). The one
274 exception is regression Model 3 (see below) in which we estimate a semiparametric
275 partially linear model. In that case, we use the the *npplr* routine in the R programming

276 language subroutine library *NP* written and maintained by Hayfield and Racine (2020),
277 and do not correct for site-specific and family-specific error correlation.

278 See the Supplemental Materials for more detailed description of the ABCD database,
279 our sample selection procedure, and the construction of the variables that we use.

280 5. Results

281 In this section, we apply our admixture regression techniques to neuropsychological
282 performance using the ABCD database. Table 1 shows means and standard deviations
283 of the regression dependent variable on data subsets sorted by SIRE choice. On the
284 full sample, by construction, the dependent variable has a mean of zero and standard
285 deviation of one. There is considerable dispersion in the subsample means sorted by
286 SIRE; for example, the means differ by 1.02 standard deviations (using the full-sample
287 standard deviation for simplicity) between two of the largest SIRE categories shown,
288 White-only SIRE and Black-only SIRE. The considerable variation in means for SIRE-
289 based subsamples provides an initial justification for performing admixture regression
290 analysis. This is a table of descriptive statistics; the standard errors shown are not
291 appropriate for formal hypothesis testing since there is no adjustment for potential
292 site-linked and family-linked correlations, particularly relevant in the case of the smaller
293 subsample categories.

294 * TABLE 1 HERE *

295 Table 2 displays empirical results from three specifications of the admixture regres-
296 sion methodology. Recall that one SIRE variable and one ancestry proportion variable
297 must be left out as an identification condition of the admixture regression: we leave out
298 the White SIRE variable and the European ancestry proportion variable. Model 1 uses
299 a linear regression specification and singleton SIRE categories for the group-identity
300 variables G ; individuals who choose multiple SIRE categories have G exposures equally
301 divided between the chosen SIRE categories as in (10). Three of the four ancestral pro-
302 portion variables and one of the six group-identity variables have statistically significant
303 coefficients. Model 2 adds a selected set of multiple-SIRE composite categories to the
304 G specification. We include the seven two-category choices with the largest number
305 of observations in our sample. Individuals with one of these two-category choices has
306 unit exposure to the associated explanatory variable, and no exposure to the weighted
307 single-SIRE variables (see equation 11 above). The same three of four ancestral pro-
308 portion variables as in Model 1 are significant in Model 2, with similar coefficients to
309 Model 1. None of the single-SIRE group identity variables is significant. Three of the
310 seven selected two-SIRE group identity variables have significantly different coefficients
311 from that implied by equal weightings of the component single-category coefficients.
312 One of these (Hispanic-Other) has a statistically significant coefficient; the other two
313 are not significantly different from zero, but are significantly different from the value
314 implied by the composite single-category coefficients. Random effects are included in
315 all models except Model 3 to capture any common variation associated with the 22
316 individual data collection sites in the ABCD study or associated with those families
317 having multiple individuals in the sample. We use the *lmer* maximum likelihood mixed
318 effects model estimation routine from the *R* language library, see Bates et al. (2015),
319 for all models except Model 3. See Nakagawa and Schielzeth (2013) for the definition
320 and interpretation of conditional and marginal R^2 in a linear mixed effects model. The
321 marginal R^2 (which does not include the explanatory power associated with site-specific
322 and family-specific random effects) is approximately 0.16 in both model specifications.
323 The African, Amerindian, and East Asian proportional ancestry variables have strong
324 and significant explanatory power in Models 1 and 2. For single-SIRE individuals, the
325 SIRE-based group identity variables are mostly indistinguishable from zero, but some of
326 the multiple-SIRE group variables are significantly different from zero.

327 * TABLE 2 HERE *

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

340 * TABLE 3 HERE *

341 * FIGURE 1 HERE *

342 * FIGURE 2 HERE *

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

353 * FIGURE 3 HERE *

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:

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

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

360 * TABLE 4 HERE *

361 Table 5 adds two new variables, US born child and Social-Economic Status (SES),
362 to the admixture regression model. US born child equals one if the child was born in
363 the USA and zero if born elsewhere. SES is a factor-analytic composite of underlying
364 variables from the ABCD database including neighborhood SES, subjective SES as de-
365 termined from a set of questionnaire answers by the parent(s)/guardian(s) of the child
366 on parental/guardian marital status, completed level of parental/guardian education,
367 reported neighborhood safety, and parental/guardian employment. See the Supple-
368 mental Materials for more detailed discussion. Models 6 and 7 are identical to Models

369 4 and 5 (respectively) from Table 3, except for the addition of these two variables. As
370 discussed in Section 3 above, including additional explanatory variables complicates the
371 interpretation of an admixture regression model in terms of the implied decomposition
372 of trait variation into linear components linked to group identities and components
373 linked to genetic ancestries. The SES variable covaries strongly with both genetic and
374 environmental components of neuropsychological performance scores. To retain the
375 standard interpretability of the admixture regression it is important to orthogonalize SES
376 with respect to the group identity and ancestry variables before running the regression.
377 For completeness, Models 6 and 7 are shown with and without the orthogonalization of
378 SES (versions a and b of each model). If the purpose of the estimation is to identify the
379 total impact of SES on the trait, the regression with raw SES is more appropriate (version
380 a). For admixture analysis intended to capture the total effects of group identity and
381 genetic ancestry on the trait, orthogonalized SES is more appropriate (version b).

382 Adding SES to the admixture regression model increases the marginal R^2 from
383 approximately 0.16 to 0.22. If SES is used in its raw form, the coefficients associated with
384 proportional ancestries tend to decrease in magnitude, but the coefficients on African,
385 Amerindian, and East Asian proportional ancestry remain strong and significant. There
386 is no clear and reliable impact on the SIRE-based group-identity coefficients from using
387 SES, in either its raw or orthogonalized form.

388 * TABLE 5 HERE *

389 6. Discussion and Limitations

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

400 This paper provides a statistical framework for admixture regression based on
401 the linear polygenic index model of behavioral genetics, and develops refinements
402 and extensions of the methodology within this framework. We provide a simple new
403 test procedure for determining whether multiple-SIRE categories have independent
404 explanatory power not captured by the individual component categories. We consider
405 additional explanatory variable in the admixture regression and their interpretation with
406 and without orthogonalization with respect to core variables. We weaken the linearity
407 assumption and develop a partially linear semiparametric regression specification.

408 We apply our methodology to neuropsychological performance test data from the
409 Adolescent Brain Cognitive Development database. We confirm existing findings that
410 genetic variation plays a role in neuropsychological performance differences across
411 self-identified races (Lasker et al. 2019). We find mixed evidence regarding the indepen-
412 dent explanatory power of multi-racial identities relative to their component single-race
413 categories. We find that when social economic status (SES) is included as an explana-
414 tory variable in the admixture regression, pre-regression orthogonalization of SES has a
415 substantial impact on the measured magnitude of the ancestry proportion coefficients.
416 We find that the proportional ancestry variable associated with African ancestry shows
417 some evidence of nonlinearity in its impact on neuropsychological performance.

418 The techniques that we propose for admixture regression studies have broad appli-
419 cability, but they do have some limitations. We describe three approaches to accommo-
420 dating multiple-identity individuals in admixture regression studies (equal weighting,
421 adding new groups, deletion of some observations) but none of the three methods is

422 fool-proof in terms of correctly capturing identity-related environmental influences in a
423 parsimonious way. We describe how to orthogonalize additional explanatory variables
424 in order to accommodate them in an admixture regression while still capturing the full
425 effect of ancestry-related genetic variation in the ancestry proportions coefficients. A
426 limitation of this orthogonalization procedure is that it does not fundamentally alter the
427 underlying regression being estimated, it merely rotates the estimated coefficients to aid
428 in their interpretation. The partially linear admixture regression method that we describe
429 has the usual limitations of nonparametric and semi-parametric estimation methods. It
430 cannot be applied with complete generality due to the curse of dimensionality, and is
431 data-intensive due to the nonparametric estimation component of the procedure.

432

433 **Funding:** This research received no external funding.

434 **Data Availability Statement:** All data used in this study comes from the Adolescent Behavior
435 Cognitive Development (ABCD) database. Qualified researchers can request access to the ABCD
436 database by applying through the National Institute of Mental Health Data Archive.

437 **Conflicts of Interest:** The authors declare no conflict of interest.

438 Appendix A

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

441 We assume that the $(k + m - 1)$ -vector of observations (s_i, G_{ij}, A_{ih}) $j = 2, \dots, k; h =$
442 $2, \dots, m$ has an i.i.d. distribution over observations $i = 1, \dots, n$ and that the conditional
443 mean functions $E[G_{ij}|A_{i2}]$ and $E[A_{ih}|A_{i2}]$ are twice differentiable throughout the interior
444 of the domain of A_2 , the closed unit interval. Let $m(\bullet)$ denote any of these conditional
445 mean functions or their first or second derivative functions. As in Racine and Li, we
446 impose the following Lipschitz-type smoothness condition on these conditional mean
447 functions and their first and second derivatives: $|m(A_2) - m(A'_2)| \leq H(z)|A_2 - A'_2|$
448 where $H(\bullet)$ is some continuous function such that $E[H(A_2)^2]$ is finite. The expectation
449 of $H(A_2)^2$ is over the probability distribution of A_2 .

450 We continue to assume that ε_i is mean-zero normally distributed with constant
451 variance. Since G_{ij} only takes the values of zero and one and A_{ih} is confined to the unit
452 interval, it necessarily follows that both have bounded fourth moments. We assume that
453 $k(\bullet)$ is a bounded second-order kernel.

454 To formally derive the limiting distribution of the Robinson estimator, it is necessary
455 to define a trimming parameter which ensures that the estimates $\widehat{\text{Pr}}(A_{i2})$ are bounded
456 away from zero. Let t denote a trimming parameter and consider the estimator described
457 in the text but where observations such that $\widehat{\text{Pr}}(A_{i2}) < t$ in (19) are dropped from the
458 subsequent estimation steps. Let ϕ denote the kernel bandwidth for sample size n .
459 Assume that the trimming parameter obeys the following two limiting conditions as
460 $n \rightarrow \infty$: $n\phi^2 t^4 \rightarrow \infty$ and $nt^{-4}\phi^8 \rightarrow 0$.

Under these conditions we have from (Robinson 1988) that

$$d \lim \sqrt{n}[(\widehat{b}_G, \widehat{b}_A) - (b_G, b_A)] \sim N(0, \sigma_\varepsilon^2 E[X'X]^{-1}).$$

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

References

1. Alexander, D.H., S.S. Shringarpure, J. Novembre and K. Lange (2015). Admixture 1.3 Software Manual, Simon Laboratory, University of Wisconsin, Bioinformatics Programs.
2. 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.

3. Cheng, C.Y., D. Reich, C.A. Haiman, A. Tandon, N. Patterson, S. Elizabeth, E.L. Akyzbekova, 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.
4. 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.
5. 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.
6. 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.
7. Halder, I., K. A. Matthews, D.J. Buysse, P.J. Strollo, V. Caser, 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.
8. Hayfield, T. and J.S. Racine (2020). Nonparametric kernel smoothing methods for mixed data types. R package np documentation.
9. 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.
10. 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.
11. 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.
12. 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.
13. 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.
14. 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.
15. 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.
16. Li, Q. and J.S. Racine (2007) *Nonparametric econometrics: theory and practice*, Princeton University Press, Princeton NJ, USA.
17. 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*.
18. 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.
19. 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.
20. 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.
21. 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.
22. Pritchard, J. K., Stephens M., and Donnelly P. (2000). Inference of population structure using multilocus genotype data. *Genetics* Vol. 155: 945–959.
23. Robinson, P.M. (1988). Root-n consistent semiparametric regression. *Econometrica*, Vol. 56: 931-954.
24. Ruderman, A., L.O. Pérez, K. Adhikari, P. Navarro, V. Ramallo, C. Gallo, R. González-José, et alia (2019). Obesity, genomic ancestry, and socioeconomic variables in Latin American mestizos. *American Journal of Human Biology*, Vol. 31: 1-13.
25. 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.
26. 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.
27. 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.
28. 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.
29. 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.

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	2.2835	-2.5520	-0.0689	0.9357	1.5102	-0.4236	-2.0061	27.0810	0.0003		
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

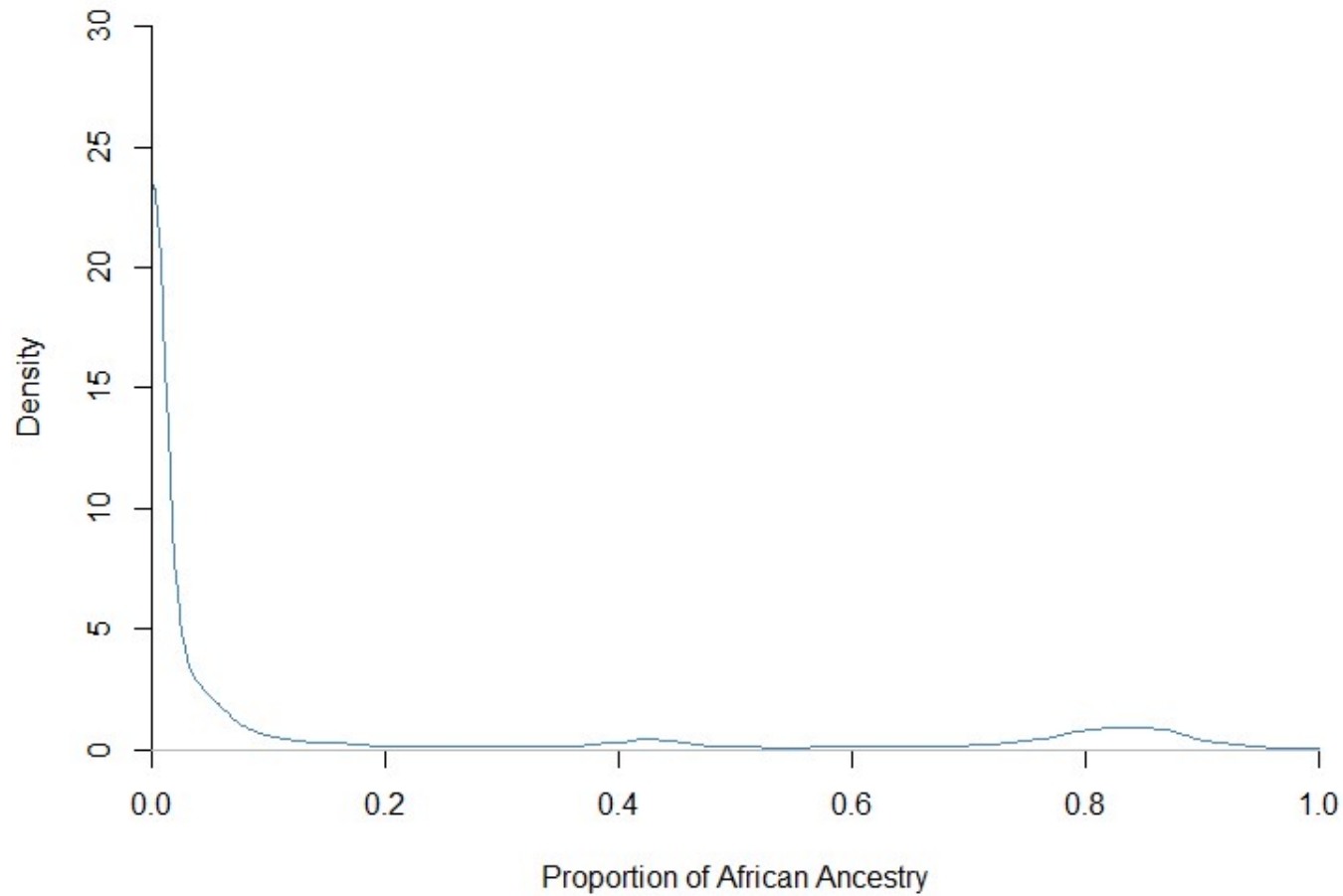


Figure 2

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

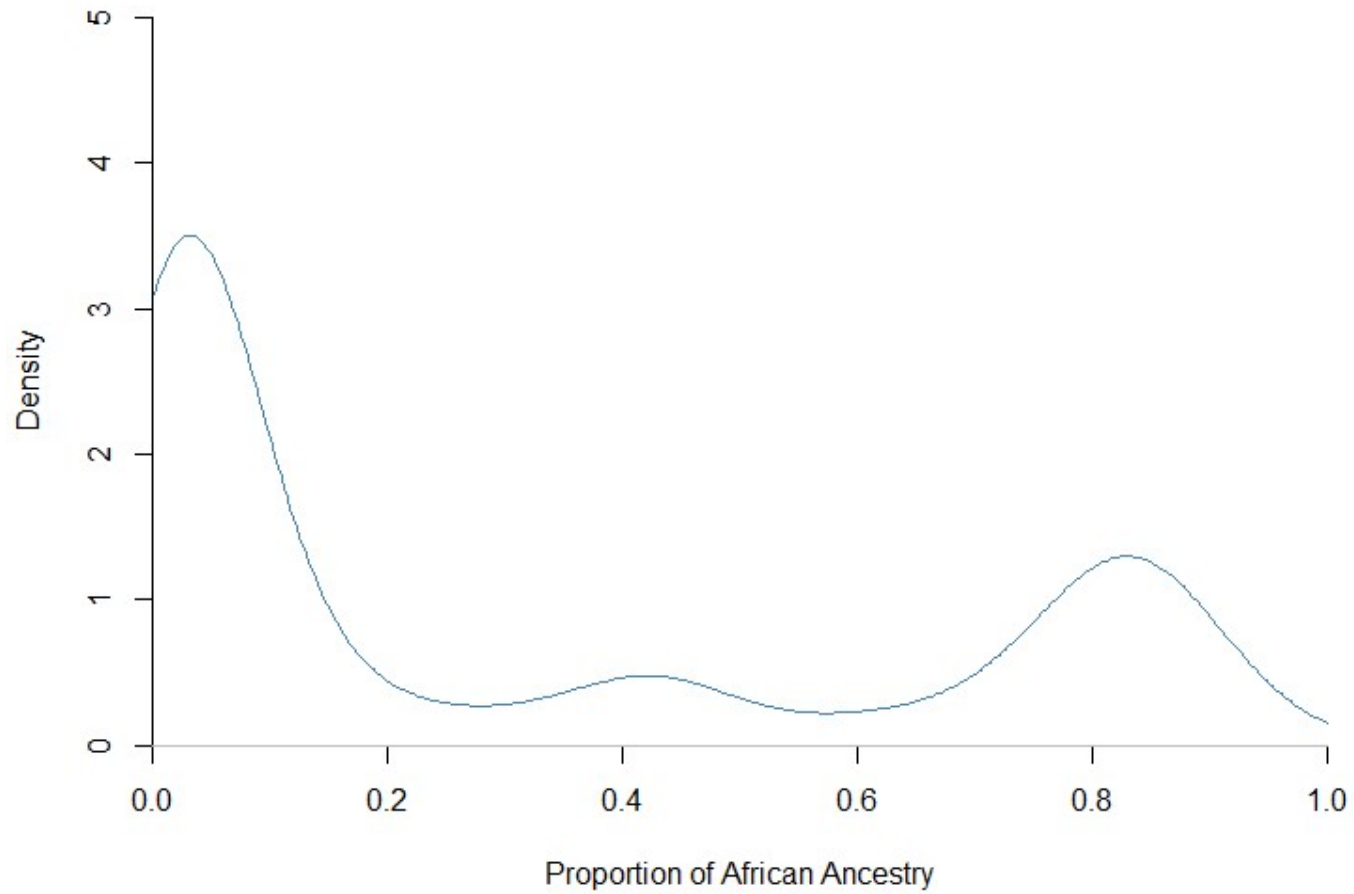


Figure 3

Linear and Nonlinear Gradients Measuring the Impact of African Ancestry

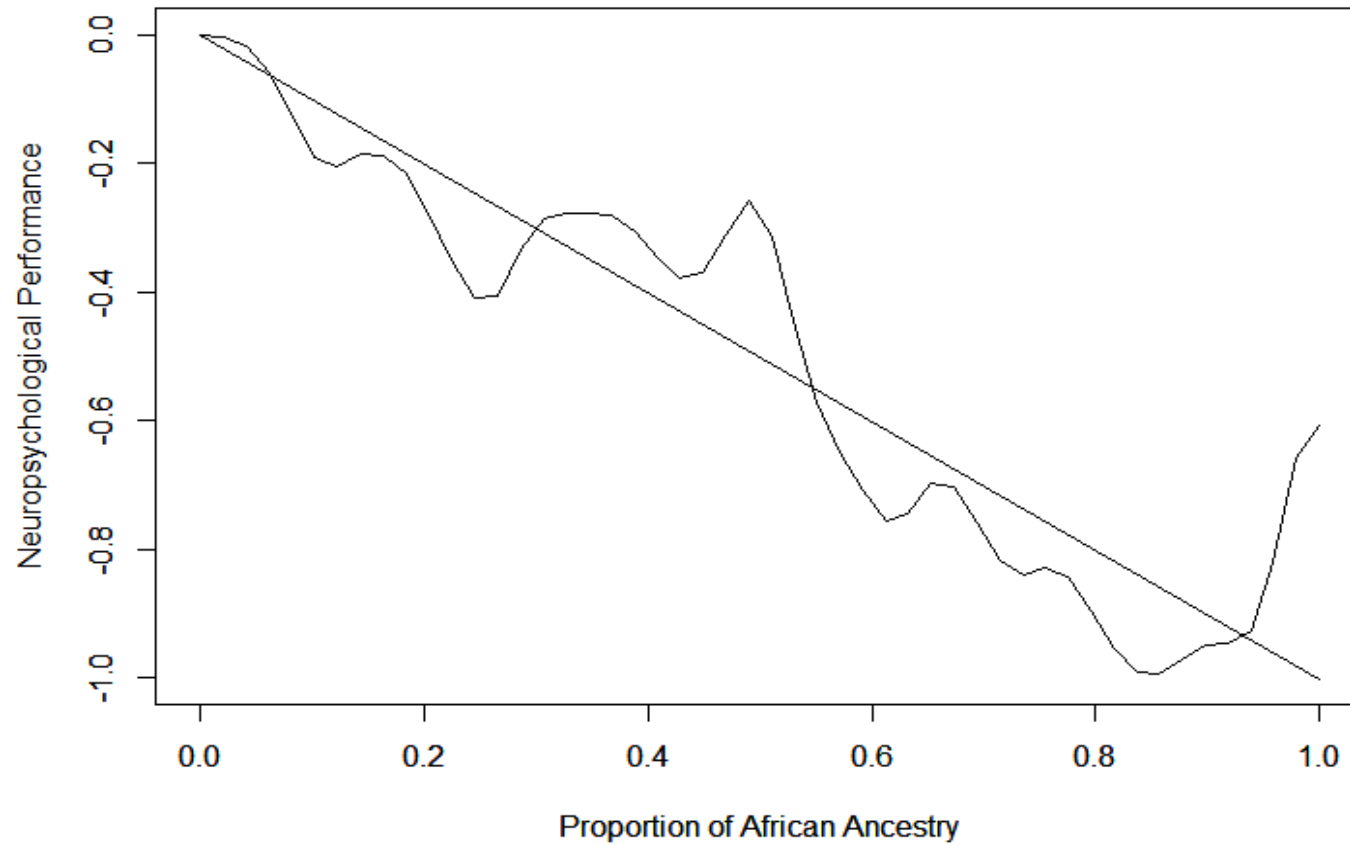


Table 4

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		2.4182	-2.5439	-0.1354	0.9303	1.5081	-0.3414	-2.0366	27.9764	0.0002	
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 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	0.1987	-1.1179	-1.3590	0.7182	0.6229	-0.1191	-0.0851	-0.2412	-0.2140	-0.1180	-0.1740
t-statistic	3.2310	-9.1130	-11.5960	3.6570	1.9630	-1.1990	-1.0430	-2.3260	-1.2530	-0.4670	-2.2900
Model 7b	0.2037	-1.1327	-1.3132	0.7316	0.6759	-0.1236	0.3584	-0.2698	-0.3505	-0.3147	-0.1593
t-statistic	3.3170	-9.0400	-10.9810	3.7120	2.1220	-1.2230	1.8800	-2.0530	-1.9730	-1.2030	-1.6040
	Piecewise Linear Variable	Socio-Economic Variables		Multiple-SIRE-Composite Explanatory Variables							
	D[A ₂ ≥0.9]A ₂	Socio-Economic Status	US Born	Black-White SIRE	Hispanic-White SIRE	Native America -White SIRE	East Asian - White SIRE	South Asian - White SIRE	Hispanic - Black SIRE	Hispanic - Other SIRE	Wald test and p-value
Model 6b [cont.]	0.2055	0.2800	0.1005								
t-statistic	2.3920	23.7280	1.7650								
Model 7b [cont.]	0.2267	0.2796	0.0897	0.0735	-0.0796	-0.1618	-0.0036	0.3184	0.0807	-0.1492	
t-statistic	2.6310	23.7100	1.5730	0.9990	-1.8180	-1.9790	-0.0360	1.7610	0.6710	-2.2240	30.5889
Test 2				2.4780	-2.7217	-0.2665	0.9170	1.5735	-0.2750	-2.0961	0.0001
Conditional R2	Models 6a,b: 0.546; Models 7a,b: 0.546										
Marginal R2	Models 6a,b: 0.218; Models 7a,b: 0.220										

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.