Online Appendix
Sibling Differences in Genetic Propensity for Education: How do Parents React?

Anna Sanz-de-Galdeano[*], Anastasia Terskaya[†]

March 13, 2023

Contents

[*] FAE, Universidad de Alicante, Carretera de San Vicente s/n, 03080 San Vicente - Alicante, Spain. Email: anna.sanzdegaldeano@gmail.com.

[†] **Corresponding author**. Department of Economics and Institut d'Economia de Barcelona (IEB), University of Barcelona, Carrer de John Maynard Keynes, 1, 11, 08034, Barcelona Spain. Email: a.terskaya@ub.edu.

A Theoretical Appendix

Solving the utility maximization problem (1) subject to (2) and (3) yields the following first order conditions:

$$\begin{gathered}
\tfrac{1}{\rho}\left\{V_1(e_1,PI_1)^{\rho}+V_2(e_2,PI_2)^{\rho}\right\}^{\frac{1}{\rho}-1}\rho\alpha_p e_1^{\rho\alpha_e}PI_1^{\rho\alpha_p-1}=\lambda p_1\\
\tfrac{1}{\rho}\left\{V_1(e_1,PI_1)^{\rho}+V_2(e_2,PI_2)^{\rho}\right\}^{\frac{1}{\rho}-1}\rho\alpha_p e_2^{\rho\alpha_e}PI_2^{\rho\alpha_p-1}=\lambda p_2\\
PI_2=\tfrac{I-p_1PI_1}{p_2}
\end{gathered}\tag{A.1}$$

, where λ is the Lagrange multiplier.

These conditions in turn yield the following expression:

$$\frac{PI_2}{PI_1}=\left\{\frac{p_1}{p_2}\left(\frac{e_2}{e_1}\right)^{\rho\alpha_e}\right\}^{\frac{1}{1-\rho\alpha_p}}=\frac{I}{p_2PI_1}-\frac{p_1}{p_2}\tag{A.2}$$

Solving (A.2) for PI_1 yields:

$$PI_1^*=\frac{I\gamma}{p_1}\tag{A.3}$$

, where $\gamma=\frac{1}{\left\{\left(\frac{p_1}{p_2}\right)^{\alpha p}\left(\frac{e_2}{e_1}\right)^{\alpha e}\right\}^{\frac{\rho}{1-\alpha p\rho}}+1}$.

Taking logs of (A.3) we obtain the following function for parental investments:

$$log(PI_1)=log(I)+G(e_1)+F\left(\frac{e_1}{e_2}\right)\tag{A.4}$$

, where $G(e_1)=-log(p(e_1))$ and $F(\frac{e_1}{e_2})=log(\gamma)$. Given that $p_i=p(e_i)$ is assumed to be a non-increasing homogeneous function of e_i of degree one, $\frac{p_1}{p_2}$ can be expressed as a function of $\frac{e_1}{e_2}$. Therefore, γ can be expressed as a function of the parameters of the model and of $\frac{e_1}{e_2}$.

Let us specify:

$$\gamma=\frac{1}{f\left(\frac{e_1}{e_2}\right)^{\frac{-\rho}{1-\alpha p\rho}}+1}\text{, where } f\left(\frac{e_1}{e_2}\right)=\left(\frac{p(e_2)}{p(e_1)}\right)^{\alpha_p}\left(\frac{e_1}{e_2}\right)^{\alpha_e}\tag{A.5}$$

Since e_1,e_2,p_1,p_2 are positive, $f\left(\frac{e_1}{e_2}\right)>0$. Also, given that α_e and α_p are positive, and $\frac{\partial p(e)}{\partial e}\leq 0$, it follows that $\frac{\partial f\left(\frac{e_1}{e_2}\right)}{\partial\frac{e_1}{e_2}}>0$.

Since we are interested in the sign of the effect of children's relative genetic endowments ($\frac{e_1}{e_2}$) on parental inputs in child 1 (PI_1) (holding constant his/her own genetic endowment level and p_1), and on how it depends on parental inequality aversion (ρ), we can obtain the sign of

$\frac{\partial log(PI_1)}{\partial f\left(\frac{e_1}{e_2}\right)}$, which is the same as the sign of $\frac{\partial log(PI_1)}{\partial \frac{e_1}{e_2}} = \frac{\partial log(PI_1)}{\partial f\left(\frac{e_1}{e_2}\right)} \underbrace{\frac{\partial f\left(\frac{e_1}{e_2}\right)}{\partial \frac{e_1}{e_2}}}_{>0}$. This yields the following expression:

$$\frac{\partial log(PI_1)}{\partial f\left(\frac{e_1}{e_2}\right)} = \gamma f\left(\frac{e_1}{e_2}\right)^{\frac{-\rho}{1-\alpha_p\rho}-1} \frac{\rho}{1-\alpha_p\rho} \tag{A.6}$$

Given that γ, $f\left(\frac{e_1}{e_2}\right)$ and $(1-\alpha_p\rho)$ are always positive, the sign of this partial effect only depends on the level of parental inequality aversion ρ. Specifically:

- $\frac{\partial log(PI_1)}{\partial f\left(\frac{e_1}{e_2}\right)} < 0$ if and only if $\rho < 0$
- $\frac{\partial log(PI_1)}{\partial f\left(\frac{e_1}{e_2}\right)} > 0$ if and only if $0 < \rho < 1$
- $\frac{\partial log(PI_1)}{\partial f\left(\frac{e_1}{e_2}\right)} = 0$ if and only if $\rho = 0$

A.1 Model Extension: Public Goods and Spillovers in Investments

We introduce sibling spillovers in parental investments (or a public good component of parental investments) by assuming that a child's human capital is a function of the parental investments he/she and his/her sibling receive following Terskaya (2023):

$$\widehat{V}_i(e_i, PI_i, PI_{-i}) = e_i^{\alpha_e}\left(PI_i^{1-\delta} PI_{-i}^{\delta}\right)^{\alpha_p} \tag{A.7}$$

, where $0 \le \delta < 0.5$ captures the degree of the public good dimension. When $\delta = 0$, parental inputs are completely separable between siblings, and the parental utility maximization problem is identical to the one previously solved. When $\delta = 0.5$, children share parental inputs equally.

The objective of the parental resource allocation problem is to maximize:

$$U = \{\widehat{V}_1^{\rho} + \widehat{V}_2^{\rho}\}^{\frac{1}{\rho}} \tag{A.8}$$

subject to (A.7) and (3).

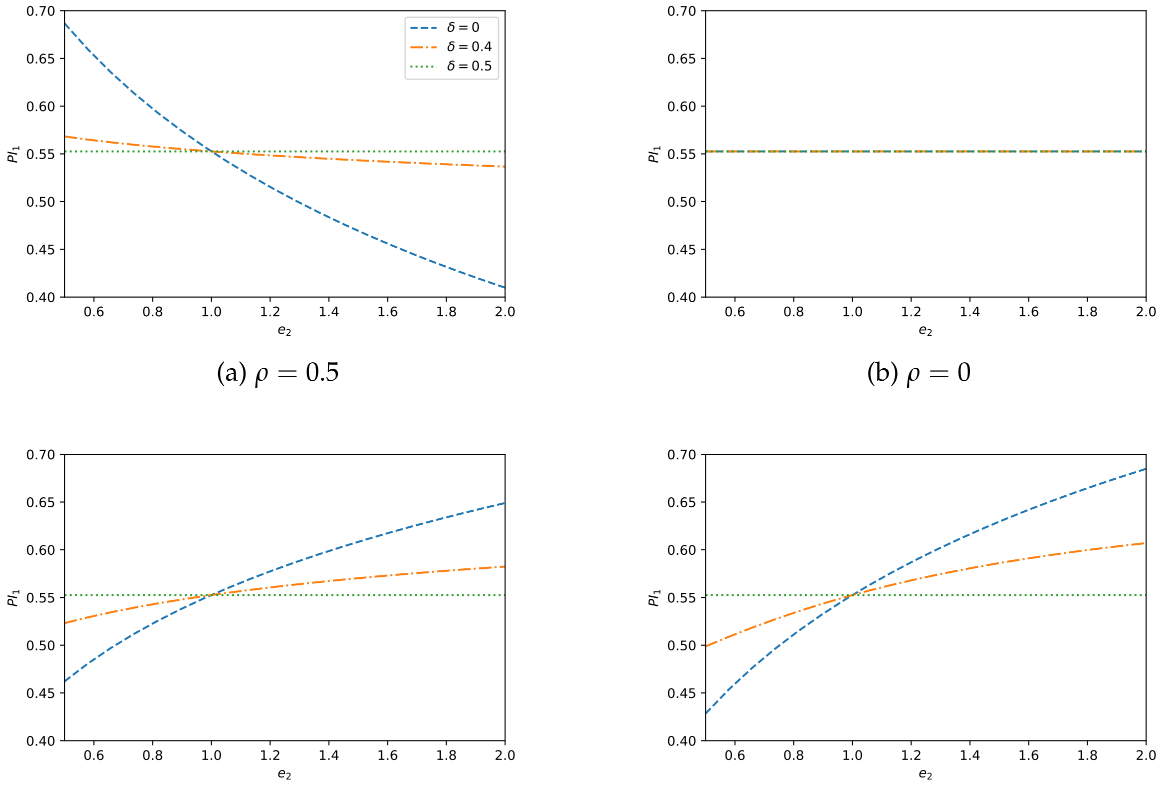
Solving the first order condition of this maximization problem yields:

$$\frac{PI_2\left((1-\delta)\hat{V}_1^{\rho} + \delta\hat{V}_2^{\rho}\right)}{PI_1\left(\delta\hat{V}_1^{\rho} + (1-\delta)\hat{V}_2^{\rho}\right)} = \frac{p_1}{p_2} \tag{A.9}$$

This expression directly implies that when $\rho = 0$, $PI_2 = \frac{p_1}{p_2} PI_1$. Substituting this into the budget constraint given by (3) yields that $PI_1 = \frac{I}{2p_1}$ and $PI_2 = \frac{I}{2p_2}$, as in the problem previously solved. Importantly, this result is the same for any value of δ. This indicates that, when parents are neutral, parental investments in child 1 are only affected by his/her own endowment (because of the price effect) but they are not affected by child 2's endowment. Importantly, the same conclusion is reached when $\delta = 0.5$ (parental investments are shared by siblings).

In Figure A.1, we analyse how PI_1 changes with e_2 using equation (A.9). The results indicate that when parents are inequality averse ($\rho < 0$) and the spillover effect is large ($\delta = 0.4$), the positive effect of child 2's endowment on parental investments in child 1 is smaller than when parental inputs are separable ($\delta = 0$). Similarly, when parents care more about efficiency ($\rho > 0$), the negative effect of child 2's endowment on parental investments in child 1 is attenuated by the sibling spillover effect.

Figure A.1: Changes in the Optimal Parental Investment in Child 1 (PI_1) with the Endowment of Child 2 (e_2)



(a) $\rho = 0.5$

(b) $\rho = 0$

(c) $\rho = -1$

(d) $\rho = -2$

Note: This Figure displays the optimal allocation of PI_1 for different values of e_2, ρ, and δ. Values are computed for $e_1 = 1$, $p_1 = e^{-0.1}$, $p_2 = e^{-0.1e_2}$, $\alpha_e = \alpha_p = 0.8$.

B Genome-Wide Association Studies

The human genome is a set of approximately 3 billion nucleotide molecules (adenine, cytosine, guanine, and thymine) in 23 chromosome pairs. [1] In approximately 99% of genome locations, there is no variation between people. The locations in the genome where there is some variation across individuals are called genetic variants or single-nucleotide polymorphisms (SNPs). For example, at a specific position, the guanine nucleotide may appear for most individuals, but for a minority of individuals this position is occupied by adenine, which means that there is a SNP at this position. The two possible molecules at a given SNP are called the major and minor alleles, where major allele refers to the molecule that is most common at a given genome position. Individual genetic information is often coded with respect to a reference genome, and each genetic variable represents the number of reference alleles (0, 1, or 2) at a given SNP.[2] There are about 10 million SNPs in the average person's genome.

Genome-wide association studies (GWAS) estimate the association between an outcome of interest (*e.g.*, educational attainment, height, body mass) and a large number of genetic variants. The analysis consists of running a set of linear regressions of an outcome on the number of reference alleles separately for each SNP. The size effects of each SNP estimated in a GWAS can be used to construct weights for summary indexes that measure individuals' genetic predisposition to different traits. These summary indexes are called polygenic indexes.[3]

In our analysis we use an educational attainment polygenic index (henceforth EA PGI) as a measure of educational genetic endowments. In the sensitivity analysis, we also use a cognitive performance polygenic index (henceforth CP PGI). To date, all GWAS of educational attainment rely on samples of individuals of European descent, so polygenic indexes are likely much less predictive and more subject to measurement error for other groups (Lee et al., 2018). Therefore, we perform our analysis on a sample of European-descent individuals, as most previous studies that use polygenic indexes.

C Genetic Data in Add Health

In Wave IV of Add Health, respondents were asked for consent for the collection of saliva samples. Approximately 80% of respondents consented to long-term archiving of their sam-

[1] A nucleotide is a molecular unit that makes up the building blocks of DNA and RNA.

[2] The reference allele often but not always coincides with the major allele.

[3] See Abdellaoui and Verweij (2021) for a comprehensive discussion of polygenic indexes and their interpretation.

ples and were eligible for genotyping. About 80% of the sample genotyping was performed with the Illumina Omni1-Quad BeadChip platform, and the Illumina Omni2.5-Quad BeadChip platform was used with the rest of the sample. After standard quality control procedures, genetic data are available for 9,974 Add Health respondents on 609,130 genetic variants (common across the genotyping platforms used).[4]

Given that the set of genotyped variants does not include all genetic variants, genetic imputation was implemented. Genetic imputation is an essential tool in the analysis of genetic associations because it increases both accuracy and precision in GWAS (Li et al., 2009). The idea of genetic imputation relies on the fact that there are groups of genetic variants that tend to always occur together because humans have common distant ancestors.Therefore, with a set of genetic markers, one can accurately infer a large number of other genetic variants. The Social Science Genetic Association Consortium (SSGAC) imputed genotypes against the Haplotype Reference Consortium $v1.1$ European reference panel using the Michigan Imputation Server. Before conducting the imputation, the sample was restricted to 5,690 European-ancestry individuals. After the imputation, the set of genetic variants was restricted to 1,211,662 $HapMap3$ genetic variants, as these variants provide good coverage of the genome in individuals of European descent and are generally well-imputed (International HapMap 3 Consortium and others, 2010).[5]

C.1 Polygenic Indexes in Add Health

In our analysis we rely on polygenic indexes for Add Health participants provided to Add Health by the Polygenic Index Repository (Becker et al., 2021). The Polygenic Index Repository uses a consistent methodology to construct polygenic indexes for 47 phenotypes in 11 datasets, including Add Health. Benjamin et al. (2021) provide a detailed guide regarding the construction and use of polygenic indexes for Add Health as part of the Polygenic Index Repository. Given the limitations of polygenic indexes for non-European ancestry individuals, polygenic indexes in the Polygenic Index Repository are only constructed in the European-ancestry subsample, leaving 5,689 individuals in Add Health with valid Repository polygenic indexes. The Add Health codebooks for the polygenic indexes produced by the Polygenic Index Repository are available online (`https://addhealth.cpc.unc.edu/wp-content/uploads/docs/restricted_use/PolygenicIndexInventories.zip`).

The Repository polygenic index of outcome t is computed as:

[4] For more information on the quality control procedures see the Add Health documentation available at: `https://addhealth.cpc.unc.edu/wp-content/uploads/docs/user_guides/AH_GWAS_QC.pdf`.

[5] For more information on imputation procedures see the Add Health documentation available at: `https://addhealth.cpc.unc.edu/wp-content/uploads/docs/user_guides/SSGAC-PGS_UsersGuide.pdf`.

$$\hat{g}_{ti} = \frac{\sum\limits_{l=1}^{L} x_{il}\hat{\gamma}_{tl}}{sd\left(\sum\limits_{l=1}^{L} x_{il}\hat{\gamma}_{tl}\right)} \tag{C.1}$$

, where x_{il} is a demeaned count of the number of reference alleles of individual i at SNP l, and $\hat{\gamma}_{tl}$ is the weight for SNP l associated with trait t. Polygenic indexes are standardized meato have mean zero and standard deviation one because each x_{il} is demeaned ($\bar{\hat{g}}_{ti} = 0$ and $sd(\hat{g}_{ti}) = 1$) and sd stands for standard deviation.

The indexes in the Repository are constructed using recent GWAS, as well as the UK Biobank and 23andMe GWAS. The approach for calculating weights for polygenic indexes in the Repository involves a Bayesian method (LDpred) that accounts for the correlation between alleles at different genome locations (Vilhjálmsson et al., 2015). Specifically, a GWAS separately estimates the effect of each genetic variant on a trait. However, some genetic variants tend to occur together (they are correlated). Hence, using the effect sizes from separate regressions estimated in a GWAS as weights for polygenic indexes without any adjustment for these correlations may limit the predictive power of polygenic indexes (Chatterjee et al., 2013). LDpred transforms the GWAS coefficients to obtain polygenic weights to account for this issue. All the Repository polygenic indexes are based on a set of approximately 1,2 million *HapMap3* SNPs.[6]

The weights for EA PGI in the Repository are constructed using the effects estimated in Lee et al. (2018) GWAS (excluding Add Health from the GWAS sample) and 23andMe GWAS for educational attainment in a sample of 1,047,538 individuals.[7] The weights for the Repository CP PGI are based on the effects estimated in Trampush et al. (2017) and the UK Biobank GWAS for cognitive performance in a sample of 260,354 individuals.

In our analysis, we define children's EA PGI using variable `pgi14` and CP PGI using variable `pgi11`.

[6] An alternative approach to construct polygenic indexes is to select a set of uncorrelated genetic variants that have the strongest association with an outcome (usually, p-value thresholds of 5×10^{-8}, 5×10^{-5}, 5×10^{-3} are used to select the variants) and to construct the polygenic indexes using only these variants and the GWAS effects as weights. However, such polygenic indexes have significantly lower predictive power than the indexes based on *HapMap3* SNPs with LDpred adjusted weights (Lee et al., 2018; Vilhjálmsson et al., 2015).

[7] It is recommended to exclude the target cohort (or close relatives of cohort members) for whom the PGI is computed (Add Health in our case) from the GWAS discovery sample in order to avoid overfitting (Wray et al., 2013; Becker et al., 2021).

C.2 Parental Polygenic Indexes

Genetic information and polygenic indexes for parents of Add Health respondent are not provided, so we construct them as follows. First, we impute genetic markers for parents using Mendelian imputation of parental genotypes, a technique proposed by Young et al. (2020). We provide further details on Mendelian imputation in Appendix D. Second, we construct parental polygenic indexes using the Polygenic Index Repository weights, which were used to construct polygenic indexes for Add Health respondents.

Since the Repository polygenic indexes are partially based on the results of the 23andMe GWAS, we first applied for an agreement with 23andMe to use their GWAS summary statistics (`https://research.23andme.com/datasetaccess/#how-to`). After obtaining this agreement, we requested non-publicly available Polygenic Index Repository weights for educational attainment that include 23andMe from the SSGAC (`https://thessgac.com/`). Since the Repository weights for cognitive performance do not include 23andMe, we downloaded publicly available Polygenic Index Repository weights for cognitive performance from the SSGAC (`https://www.thessgac.org/pgi-repository`).

Using the obtained weights and imputed parental genetic markers, we constructed parental polygenic indexes using Young et al. (2020) Python package SNIPar (fPGS.py routine). The package is publicly available at `https://github.com/AlexTISYoung/SNIPar`.

D Mendelian Imputation of Parental Genotypes

To obtain genetic information for parents of Add Health respondents, we apply Mendelian imputation of parental genotypes, a technique proposed by Young et al. (2020).

This method imputes parental genotypes taking advantage of the fact that genetic data on siblings contain information on the genotypes of parents because genes are inherited from parents. Specifically, this method relies on identity-by-descent analysis of siblings' genetic variants, using information on whether siblings inherited the same or different genetic variants from each parent.[8] Therefore, this method requires data on siblings' alleles for each SNP. For instance, suppose that sibling 1 has the $++$ variant at genome location l and sibling 2 has the $--$ variant at the same location. Hence, sibling 1 has inherited a $+$ from the mother and a $+$ from the father, and sibling 2 has inherited a $-$ from the mother and a $-$ from the father, which implies that the mother and the father have the $-+$ variant at location l. In contrast,

[8] Identity-by-descent is a term used in genetics to describe a situation where an individual inherits a particular genetic variant or mutation from one of their ancestors. This can happen when two individuals who are related to each other, such as siblings or cousins, both inherit the same genetic variant from a common ancestor. In this case, the individuals are said to be in an identity-by-descent state with respect to that particular genetic variant.

suppose that sibling 1 and 2 have the ++ variant. Then we know that both the mother and the father have at least one + allele, but we cannot identify with certainty the other alleles that parents have. In this case, the method uses the population allele frequency.

Since it is not possible to identify whether each allele was inherited from the mother or from the father, the sum of parental genotypes is imputed. Hence, we impute $x_{pfl} = x_{mfl} + x_{ffl}$, where x_{mfl} and x_{ffl} denote the maternal and paternal allele count at SNP l in family f, respectively.

Using this method, we imputed parental alleles on 1,211,662 $HapMap3$ SNPs using siblings' $HapMap3$ SNPs. We use the genetic matrices for Add Health respondents from the database of Genotypes and Phenotypes (dbGaP, `https://www.ncbi.nlm.nih.gov/gap/`). For imputation of parental genes, we use Young et al. (2020) Python package SNIPar (impute_runner routine). The package is publicly available at `https://github.com/AlexTISYoung/SNIPar`. For the imputation, we computed identity-by-descent segments using the algorithm for relationship inference in genome-wide association studies proposed by Manichaikul et al. (2010). The algorithm is implemented in a publicly available software package, KING (we use the ibdseg KING command), available for download at `https://www.kingrelatedness.com/`.

E Genetic Ability and the Genetic Component of Educational Attainment

The objective of this section is to analyze how environmental responses to genetic ability may alter the interpretation of the estimated effects of EA PGI on parental investments. Specifically, the measure of genetic ability that we use is the EA PGI, that is, a best linear genetic predictor of educational attainment. Educational attainment might be affected by genetic ability through certain cognitive and psychological characteristics and through environmental responses (e.g., parental investments). Therefore, the weights used to compute EA PGI may capture such indirect environmental effects. This may imply that in some cases, EA PGI is not a good proxy for educational genetic ability. In this Appendix we analyse this issue in greater detail.

We deviate from the theoretical model described in Section 2 and we consider that there are many potential sources of inputs or investments that may affect human capital. We focus on educational attainment as a measure of human capital, as this is the phenotype used to compute EA PGI. Furthermore, we consider different dimensions of genetic endowments, which may affect educational attainment directly and through environmental responses.

Let us denote educational attainment of individual i by EA_i. I_{ki} denotes inputs of type

$k = 0, 1, ..., K$ that affect EA_i, which include parental investments denoted by I_{0i}, and other possible inputs (e.g., inputs from teachers, peers, etc.). G_{Ai} denotes child i's overall genetic educational ability net of environmental factors, which is possibly a non-linear function of i's genetic variants and it is assumed to be standardized ($sd(G_{Ai}) = 1$). For simplicity of notation, in the remainder of this section we omit subscript i.

Assume that educational attainment is given by:[9]

$$EA = \psi G_A + \sum_k \omega_k I_k + \epsilon \tag{E.1}$$

, where ψG_A captures the direct effect (net of environmental responses) of genetic endowments on educational attainment, and ψ and ω_k are both positive constants. We are interested in estimating the effect of educational genetic ability G_A on parental investments.

As in Section 2, inputs, which include parental investments, are defined endogenously and may depend on children's initial genetic endowments. Specifically, we assume that:

$$I_k = \theta_k G_k + q_k \ \ \forall k \tag{E.2}$$

, where G_k denotes some initial genetic characteristics of a child (a function of genetic variants assumed to be standardized such that $sd(G_k) = 1$) which affect the level of inputs I_k, and q_k denotes other characteristics. For instance, for parental investments, this equation follows from (4) and implies that:

$$I_0 = \theta_0 G_0 + \theta_0^s G_0^s + u_0 \tag{E.3}$$

, where G_0^S denotes siblings' G_0, and u_0 may include family characteristics and other non-genetic factors such that $q_0 = \theta_0^s G_0^s + u_0$.

Note that G_k can be decomposed into two terms: a term correlated with G_A, and a term uncorrelated with G_A (standardized to have variance 1), such that:

$$G_k = b_k G_A + \sqrt{1 - b_k^2} G_{k\sim A} \ \ \forall k \tag{E.4}$$

, where b_k is the correlation coefficient between G_k and G_A. This implies that the effects of interest of G_A and sibling's G_A (G_A^S) on I_0 are $\theta_0 b_0$ and $\theta_0^S b_0$, respectively.

However, in practice we do not observe G_A, and therefore we cannot estimate its effect on I_0. Instead, we have a measure of the genetic component of educational attainment, G_{EA},

[9] Note that if we assume that there are only parental inputs ($\sum_k \omega_k I_k = \omega_0 I_0$), this equation is equivalent to equation (2) with $EA = log(V)$, $G_A = log(e)$, and $I_0 = log(PI)$.

which includes the effect of educational genetic ability, G_A, and the indirect effects of genes on educational attainment through environmental responses. Therefore, in our analysis we make the following assumption:

- Assumption E.1. *Genetic educational ability G_A is well proxied by the genetic component of educational attainment G_{EA}.*

The analysis below establishes the conditions under which this assumption holds and assesses the bias in the estimated effect of G_A on parental investments if it is violated.

Substituting (E.4) and (E.2) into (E.1) yields:

$$\begin{aligned} EA = \theta_{EA} G_{EA} + u_S = \psi G_A + \textstyle\sum_k \omega_k(\theta_k G_k + q_k) + \epsilon = \\ G_A(\psi + \textstyle\sum_k \omega_k \theta_k b_k) + \textstyle\sum_k \omega_k \theta_k \sqrt{1 - b_k^2} G_{k\sim A} + \textstyle\sum_k \omega_k q_k + \epsilon \end{aligned} \tag{E.5}$$

, such that $\theta_{EA} G_{EA} = G_A(\psi + \sum_k \omega_k \theta_k b_k) + \sum_k \omega_k \theta_k \sqrt{1 - b_k^2} G_{k\sim A}$ and $u_S = \sum_k \omega_k q_k + \epsilon$.

The first term of equation (E.5) indicates that EA can be both directly and indirectly affected by genetic ability G_A. The second term indicates that educational attainment is also indirectly affected by other genetic characteristics uncorrelated with G_A ($G_{k\sim A}$).

Next, we show which effects will be captured by regressing I_0 on EA PGI. However, we need to impose several assumptions regarding the functional forms of G_A, G_k, and G_{EA}:

- Assumption E.2. *Genetic characteristics (G_A and G_k $\forall k$) are proxied by additive genetic factors ($\bar{G}_A$ and $\bar{G}_k$ $\forall k$), i.e., linear combinations of genetic variants (ignoring dominance and epistasis).*[10]

  This assumption is supported by a large and growing body of research that we now discuss. While non-additive genetic factors may be potentially important, abundant evidence from quantitative genetics indicates that most genetic variance in the analysis of population data can be captured by additive genetic factors (Falconer and Mackay, 1996). For instance, Hill et al. (2008) evaluate the evidence from empirical studies of genetic variance components and conclude that additive variance accounts for over half of and often close to the total genetic variance. Interestingly, their theoretical model shows that the proportion of the additive genetic variance with respect to the total genetic variance is high even in the presence of dominance and epistasis because most

[10] *Genetic dominance* occurs when a particular gene variant, or allele, is expressed more frequently and consistently in an individual or population than other gene variants for a particular trait. This means that the dominant allele has a stronger effect on the trait than the other, non-dominant alleles. *Genetic epistasis* is the phenomenon in which one gene influences the expression of another gene. This means that the effect of one gene on a particular trait can be modified by the presence of one or more other genes.

variants have low minor allele frequency. A recent study by Hivert et al. (2021) estimates non-additive genetic variances in human complex traits using genome-wide data and finds that the average across traits dominance and epistasis genetic variances are smaller than the additive genetic variance (they estimate that average across traits additive, dominance, and epistatic genetic factors amount to 0.208, 0.001, and 0.055, respectively). Similarly, Zhu et al. (2015) estimate that the proportion of dominance genetic variance across 79 traits is approximately a fifth of the proportion of the additive genetic variance. A recent study by Palmer et al. (2021) evaluates genetic dominance effects in more than 1,000 phenotypes in the UK Biobank GWAS and finds no evidence that genetic dominance contributes to phenotypic variation. Specifically, their results indicate that additive components explain on median (across phenotypes) 21 times more of the phenotypic variance than uncorrelated non-additive components. Along the same lines, Okbay et al. (2022) provide evidence that the proportion of dominance genetic variance is negligible for educational attainment.

In what follows, we denote the additive genetic ability and the additive genetic factors relevant for input k by $\bar{G}_A$ and $\bar{G}_k$, respectively.

- Assumption E.3. *The direct-effect additive genetic factor for educational attainment is proxied by the between-family additive genetic factor for educational attainment.*

  Equation (E.1) indicates that schooling is affected by q_k, which may include parents' and siblings' genetic characteristics that are correlated with children's own genetic characteristics. Between-family GWAS used to compute polygenic index do not control for these characteristics, which may bias polygenic weights. To avoid this issue, one would need to rely on the results of a between-family GWAS that controls for parental genes. The best linear genetic predictor net of the effect of parental genes is referred to by the "direct-effect" additive genetic factor.

  In their Section "Within-Family Analyses" (p. 198), Trejo and Domingue (2018) derive the bias in the effect sizes in a within-family model based on a between-family GWAS. They conclude that the effect sizes would be deflated by a factor equal to the correlation between the direct-effect additive genetic factor (obtained in the GWAS that controls for parental genes) and the additive genetic factor (obtained in the ordinary GWAS that does not control for parental genes). Young et al. (2020) document that this correlation for education attainment is 0.739, suggesting that the estimated effects found using between-family GWAS results are likely to be deflated by approximately this factor.

The standardized additive genetic factor of educational attainment is the best linear prediction of education attainment (EA) given the genetic variants. This linear combination can be obtained from equation (E.5). Under the assumptions listed above, the additive genetic factor of educational attainment is:

$$\bar{G}_{EA} = \frac{\bar{G}_A(\psi + \sum_k \omega_k \theta_k b_k) + \sum_k \omega_k \theta_k \sqrt{1 - b_k^2} \bar{G}_{k \sim A}}{sd(\bar{G}_A(\psi + \sum_k \omega_k \theta_k b_k) + \sum_k \omega_k \theta_k \sqrt{1 - b_k^2} \bar{G}_{k \sim A})} \tag{E.6}$$

Let us consider a special case in which $G_k = G_A \; \forall \; k$. In this case, (E.6) transforms into:

$$\bar{G}_{EA} = \frac{\bar{G}_A(\psi + \sum_k \omega_k \theta_k)}{sd(\bar{G}_A(\psi + \sum_k \omega_k \theta_k)} = \begin{cases} \bar{G}_A \quad if \;\; \psi + \sum_k \omega_k \theta_k > 0 \\ -\bar{G}_A \quad if \;\; \psi + \sum_k \omega_k \theta_k < 0 \end{cases} \tag{E.7}$$

Given that $\psi > 0$ and $\omega_k > 0 \; \forall \; k$, $\psi + \sum_k \omega_k \theta_k$ can be negative only if θ_k is negative for many inputs. This would imply that there is full compensation of educational attainment and that EA is inversely related to genetic ability, which is unlikely to be the case. Therefore, we impose an additional assumption:

- Assumption E.4. *The effect of inputs does not fully compensate the effect of genetic ability on educational attainment. That is, genetic ability G_A is positively correlated with the additive genetic component of educational attainment $\bar{G}_{EA}$.*

  This assumption is supported by empirical evidence that the genetic variants included in EA PGI are positively associated with brain volume, white-matter tract integrity, and neuronal development and function (*e.g.*, Rietveld et al. 2013; Elliott et al. 2019; Demange et al. 2021; Lee et al. 2018).

Under this assumption, the additive genetic component of educational attainment is the correct regressor since $\bar{G}_A = \bar{G}_{EA}$.

When $\bar{G}_k \neq \bar{G}_A$, the magnitude of the estimated effects of $\bar{G}_A$ will generally be biased, but the signs of the estimated effects of G_A will be correct when there is not full compensation ($\psi + \sum_k \omega_k \theta_k > 0$).

In particular, it can be shown that the estimated effects of G_A on I_0 when $\bar{G}_{EA}$ is used as a proxy for G_A will be inflated/deflated by the following factor:

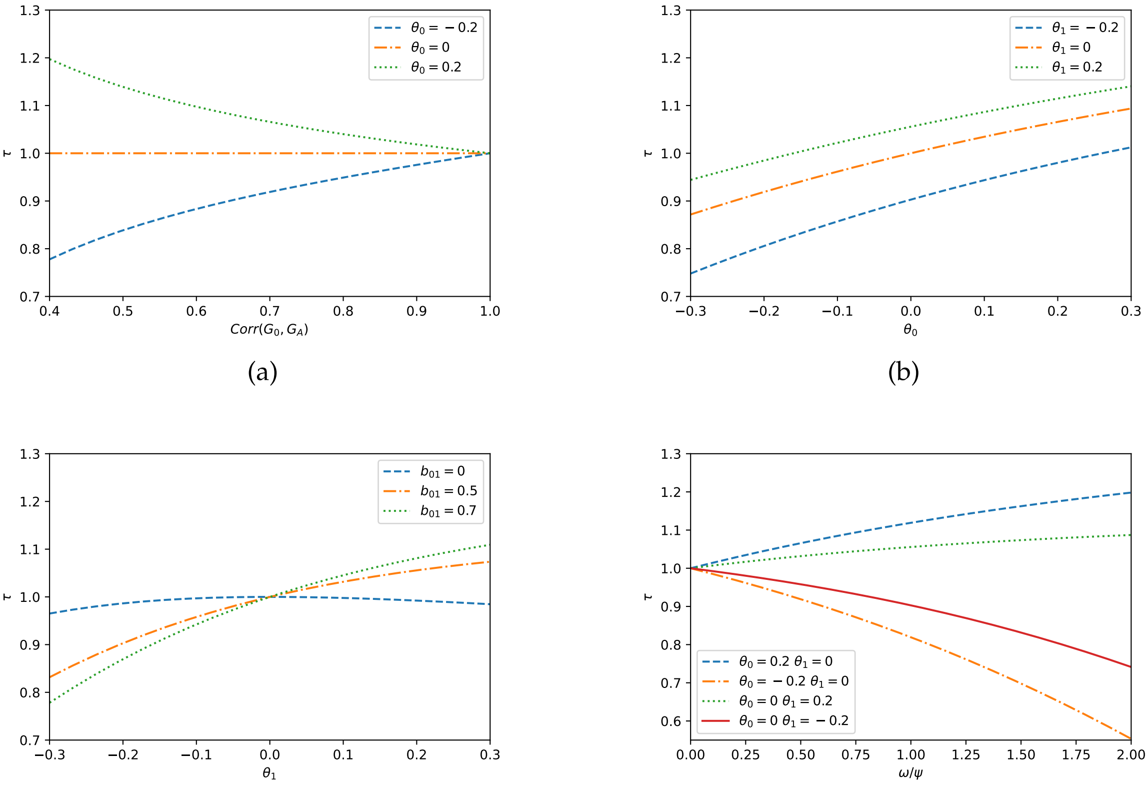
$$\tau = \frac{\psi + \sum_k \omega_k \theta_k (b_k + (1 - b_k^2) b_0^k / b_0)}{sd(\bar{G}_A(\psi + \sum_k \omega_k \theta_k b_k) + \sum_k \omega_k \theta_k \sqrt{1 - b_k^2} \bar{G}_{k \sim A})} \tag{E.8}$$

, where b_0^k is obtained by decomposing $\bar{G}_{k \sim A}$ into a correlated and an uncorrelated component with $\bar{G}_{0\sim A}$: $\bar{G}_{k\sim A} = b_0^k \bar{G}_{0\sim A} + \sqrt{1-(b_0^k)^2}\bar{G}_{k\sim A\sim 0}$. Note that $b_0^0 = 1$.

(E.8) indicates that when there is no compensation or reinforcement ($\theta_k = 0 \;\; \forall k$) or when EA is not affected by investments ($\omega_k = 0 \;\; \forall k$), the size of the coefficients will be correctly estimated ($\tau = 1$). When there is no compensation or reinforcement of parental investments ($\theta_0 = 0$), and $\bar{G}_k$ is correlated with $\bar{G}_0$ only through $\bar{G}_A$ ($b_0^k = 0 \;\; \forall k \sim 0$), the estimated effect of $\bar{G}_A$ will be deflated ($\tau < 1$) unless $\theta_k = 0 \;\; \forall k$.

In order to analyze the magnitude of the potential bias, we simulate (E.8) for the model with two inputs (I_0 and I_1) imposing different assumptions about parameter values listed in Table E.1.

Figure E.1: Relative Amount of Bias Driven by Environmental Responses to Genetic Ability



(a)

(b)

(c)

(d)

Note: The figures simulate the inflation factor τ (on the vertical axes) in the estimated effect of $\bar{G}_A$ on I_0 when $\bar{G}_{EA}$ is used as a proxy for $\bar{G}_A$ computed according to equation (E.8). A value of 1 indicates that the estimated effect of $\bar{G}_A$ is unbiased. A value of 0.8 indicates that the estimated effect will be 80% of the true effect. Panel (a) shows the association between the inflation factor and the correlation between $\bar{G}_0$ and $\bar{G}_A$ (on the horizontal axes). Panel (b) shows the association between the inflation factor and the effect of children's genetic characteristic on parental investments (θ_0) for different values of the effects of children's genetic characteristics and non-parental investments (θ_1). Panel (c) shows the association between the inflation factor τ and θ_1 assuming that $\theta_0 = 0$ for different values of the correlation between genetic characteristics that affect parental and non-parental inputs (net of true genetic ability $\bar{G}_A$), $b_0^1 = Corr(G_{0\sim A}, G_{1\sim A})$. Panel (d) shows the association between the inflation factor and the relative effect of inputs (ω_0 and ω_1) with respect to the direct effect of genetic ability on educational attainment (ψ) for different values of θ_0 and θ_1. Parameter values are listed in Table E.1.

First, we show how the bias will change with the correlation between $\bar{G}_0$ and $\bar{G}_A$ for dif-

Table E.1: Parameter Values Used for Simulations in Figure E.1

	Panel A	Panel B	Panel C	Panel D
ψ	1	1	1	1
ω_0	0.5	0.5	0.5	$0-2$
ω_1	1	1	1	$0-2$
θ_0	$-0.2, 0, 0.2$	$-0.3-0.3$	0	$-0.2, 0, 0.2$
θ_1	0	$-0.2, 0, 0.2$	$-0.3-0.3$	$-0.2, 0, 0.2$
$b_0 = Corr(\bar{G}_0, \bar{G}_A)$	$0.4-1$	0.7	0.7	0.7
$b_1 = Corr(\bar{G}_1, \bar{G}_A)$	0.7	0.7	0.7	0.7
$b_0^1 = Corr(\bar{G}_{0\sim A}, \bar{G}_{1\sim A})$	0.5	0.5	$0, 0.5, 0.7$	0.5

ferent values of θ_0, while assuming that $\theta_1 = 0$ (non-parental investments are neutral). Panel (a) of Figure E.1 shows that when there is compensation ($\theta_0 < 0$) or reinforcement ($\theta_0 > 0$) of parental investments, the estimated effects of $\bar{G}_{EA}$ will underestimate and overestimate the effect of $\bar{G}_A$ on parental investments, respectively, but this bias is small when the correlation between $\bar{G}_0$ and $\bar{G}_A$ is sufficiently large.

In our second experiment we analyze how the bias will change depending on θ_0 and assuming several values of θ_1. Panel (b) of Figure E.1 shows that the effect can be unbiased when there is compensation of some inputs but reinforcement of other inputs. When inputs are highly compensatory, the estimated effect of $\bar{G}_{EA}$ will underestimate the effect of $\bar{G}_A$, while when inputs are reinforcing, the effect will be overestimated. For instance, when non-parental inputs are as important for schooling as genetic ability and parental inputs are half as important ($\phi = 2\omega_0 = \omega_1$), $Corr(\bar{G}_0, \bar{G}_A) = Corr(\bar{G}_1, \bar{G}_A) = 0.7$, $Corr(\bar{G}_{0\sim A}, \bar{G}_{1\sim A}) = 0.5$, and both inputs are compensatory ($\theta_0 = \theta_1 = -0.2$), the estimated effect of $\bar{G}_{EA}$ will account for 80.6% of the true effect of $\bar{G}_A$. When both inputs are reinforcing ($\theta_0 = \theta_1 = 0.2$), the estimated effect of $\bar{G}_{EA}$ will account for 111.5% of the true effect of $\bar{G}_A$. When parental inputs are compensatory but non-parental inputs are reinforcing ($\theta_0 = -0.2, \theta_1 = 0.2$), the estimated effect of $\bar{G}_{EA}$ will account for 98.5% of the true effect of $\bar{G}_A$.

In our third experiment we analyze how the bias will change for different values of θ_1, while assuming that $\theta_0 = 0$ (parental inputs are neutral). This experiment is motivated by our results that the overall effect of children's ability on parental investments is not significant and close to zero ($\hat{\beta}_1 + \hat{\beta}_2 \approx 0$). Panel (c) of Figure E.1 depicts the values of the bias for different values of $Corr(\bar{G}_{0\sim A}, \bar{G}_{1\sim A})$ and θ_1. The results indicate that, even when there is strong compensation or reinforcement of non-parental inputs, the bias in the effect of $\bar{G}_A$ on I_0 is small. For instance, when there is reinforcement of non-parental inputs ($\theta_1 = 0.2$), again assuming that non-parental inputs are as important for schooling as genetic ability, that parental inputs are half as important, $Corr(\bar{G}_A, \bar{G}_0) = Corr(\bar{G}_A, \bar{G}_1) = 0.7$, and $Corr(\bar{G}_{0\sim A}, \bar{G}_{1\sim A}) = 0.5$, our estimates will account for 105.6% of the true effect of $\bar{G}_A$. If there is compensation ($\theta_1 = -0.2$), the estimated effect of $\bar{G}_{EA}$ will account for 90.3% of the true effect of $\bar{G}_A$.

Finally, we look at how the bias will change depending on the relative importance of the direct effect of genetic ability (ψ) and the effects of inputs (ω_0, ω_1). Panel D of Figure E.1 depicts the size of the bias depending on $\omega_0/\psi = \omega_1/\psi$ for different values of θ_0 and θ_1. The results indicate that the bias will be small if inputs have relatively low importance for schooling with respect to the genetic ability. For example, if there is compensation of non-parental inputs ($\theta_1 = -0.2$) and the effects of both parental and non-parental inputs constitute 50% of the direct genetic effect (assuming again that $Corr(\bar{G}_A, \bar{G}_0) = Corr(\bar{G}_A, \bar{G}_1) = 0.7$, and $Corr(\bar{G}_{0\sim A}, \bar{G}_{1\sim A}) = 0.5$), our estimates will account for 95.8% of the true effect of $\bar{G}_A$, and if there is reinforcement ($\theta_1 = 0.2$), they will account for 103.2% of the true effect. If parental and non-parental inputs are twice as important as the direct effect of genetic ability, our estimated effects will account for 74.1% and 108.7% of the true effects if there is compensation or reinforcement, respectively. Note that our measure of parental investments increases educational attainment by about 0.11 standard deviations, which is about 3 times smaller than the effect of EA PGI on educational attainment (see Tables 2 and H.4).

In sum, our results suggest that compensatory inputs (parental and non-parental) may lead to underestimating the effect of genetic ability on parental investments, while reinforcing inputs may lead to overestimating this effect. The bias will be negligible when inputs (parental or non-parental) are mainly affected by the same genetic characteristics that directly affect educational attainment, or when parents display neither reinforcing nor compensating behaviors (which is consistent with our estimates). Moreover, the bias will be negligible when the direct effect of genetic ability on schooling is relatively important with respect to the effect of inputs.

F Measurement Error in the Polygenic Index

In this Appendix we discuss the measurement error problem that arises when using EA PGI as a measure of the additive genetic component of educational attainment. Since the weights for the genetic variants used to compute the EA PGI are unobserved and need to be estimated, the EA PGI is a noisy measure of the additive genetic component of educational attainment. Moreover, the EA PGI is only based on measured genetic variants, whereas the additive genetic factor it is used to proxy is based on all genetic variants. Below we provide a theoretical framework used to adjust the estimated effect sizes for this measurement error. The framework is built upon Becker et al. (2021), who derive a measurement error correction method for regressions of a phenotype on a PGI. We derive a similar method for regressions of a phenotype on individuals' own PGI, siblings' PGI, and parental PGI.

F.1 Setup

Consider a phenotype y^*. The allele count for individual i for genetic variant j is denoted by x^*_{ij}. In the derivations we use a mean-centered transformation, so that $y_i = y^*_i - \mathbb{E}(y^*_i)$ and $x_{ij} = x^*_{ij} - \mathbb{E}(x^*_{ij})$. Denote the vector of mean-centered allele counts across J observed SNPs by $X_i = (x_{i1}, x_{i2}, ..., x_{iJ})$ and the vector of mean-centered allele counts across K unobserved SNPs by $Z_i = (x_{iJ+1}, x_{iJ+2}, ..., x_{iJ+K})$. The best linear prediction of y_i given by *all* SNPs is:

$$\mathbf{G}_i = \frac{X_i\gamma + Z_i\delta}{sd(X_i\gamma + Z_i\delta)} \tag{F.1}$$

, where γ and δ minimize $\mathbb{E}\Big[(y_i - X_i\gamma - Z_i\delta)^2\Big]$, $\mathbf{G}_i$ is referred to as the *standardized additive genetic factor*, and the proportion of the variance of y_i explained by the standardized additive genetic factor is called *narrow-sense heritability* (h^2).

The best linear prediction of y_i given by J observed SNPs is:

$$g_i = \frac{X_i\gamma}{sd(X_i\gamma)} \tag{F.2}$$

, where γ minimizes $\mathbb{E}\Big[(y_i - X_i\gamma)^2\Big]$, g_i is referred to as the *standardized additive SNP factor,* and the proportion of the variance of y_i explained by g_i is referred to as *SNP heritability* (h^2_{SNP}).

Similarly, we can define the best linear prediction of y_i given by unobserved K SNPs as $g^{miss}_i = \frac{Z_i\delta}{sd(Z_i\delta)}$, where g_i is referred to as the *missing additive genetic factor.*

Assuming that $Corr(g^{miss}_i, g_i) = 0$,

$$\mathbf{G}_i = \sqrt{\pi}g_i + \sqrt{1-\pi}g^{miss}_i \tag{F.3}$$

, where $\pi = \frac{Var(X_i\gamma)}{Var(X_i\gamma)+Var(Z_i\delta)} = \frac{h^2_{SNP}}{h^2} \leq 1$, as $h^2_{SNP} = \frac{Var(X_i\gamma)}{Var(y_i)}$ and $h^2 = \frac{Var(X_i\gamma+Z_i\delta)}{Var(y_i)}$.

F.2 The standardized additive SNP factor as a proxy for the standardized additive genetic factor

Consider the model

$$Y_i = \beta_{\mathbf{G}}\mathbf{G}_i + \beta_{\mathbf{G}_s}\mathbf{G}_{si} + \beta_{\mathbf{G}_p}\mathbf{G}_{pi} + z'_i\delta_{\mathbf{G}} + \varepsilon_i \tag{F.4}$$

, where $\mathbf{G}_i$, $\mathbf{G}_{si}$ and, $\mathbf{G}_{pi}$ are standardized additive genetic factors of i, i's sibling, and i's parents, respectively, and z_i denotes i's characteristics. This model is identical to (5), where $\beta_1 = -\beta_{\mathbf{G}_s}$, $\beta_2 = \beta_{\mathbf{G}} + \beta_{\mathbf{G}_s}$.

Now suppose that we use q_i, q_{si}, and q_{pi} as proxies for $\mathbf{G}_i$, $\mathbf{G}_{si}$, and $\mathbf{G}_{pi}$ respectively. Then, equation (F.4) can be rewritten as:

$$Y_i = \beta_g g_i + \beta_s g_{si} + \beta_p g_{pi} + z_i' \delta_g + u_i \tag{F.5}$$

, where

$$\begin{aligned} \beta_g &= \sqrt{\pi}\beta_{\mathbf{G}} \\ \beta_s &= \sqrt{\pi}\beta_{\mathbf{G}_s} \\ \beta_p &= \sqrt{\pi}\beta_{\mathbf{G}_p} \\ u_i &= \varepsilon + \sqrt{1-\pi}\Big(\beta_{\mathbf{G}} g_i^{miss} + \beta_{\mathbf{G}_s} g_{si}^{miss} + \beta_{\mathbf{G}_p} g_{pi}^{miss}\big) \end{aligned} \tag{F.6}$$

Given that the composite error term u_i is uncorrelated with g_i, g_{si}, and g_{pi}, the OLS regression of Y_i on g_i, g_{si}, and g_{pi} will produce consistent estimators of β_g, β_s, and β_p. Note however that (F.6) implies that the effects of one standard deviation increases in the standardized additive SNP factors (g_i, g_{si}, g_{pi}) are $\sqrt{\pi} = \sqrt{\frac{h^2_{SNP}}{h^2}} \leq 1$ times the effects of one standard deviation increases in the standardized additive genetic factors ($\mathbf{G}_i$, $\mathbf{G}_{si}$, $\mathbf{G}_{pi}$).

Note that if narrow-sense heritability h^2 is two times larger than SNP heritability h^2_{SNP}, as it is suggested by Cheesman et al. (2017), the effect of one standard deviation increase in the additive SNP factor is $\sqrt{2}$ times smaller than the effect of one standard deviation increase in the additive genetic factor. We use the former interpretation and we use model (F.5) to estimate $\beta_g, \beta_s, \beta_p$, while acknowledging that the effects of a standard deviation increase in the additive genetic factors are about $\sqrt{2}$ times larger.

F.3 Measurement error in the standardized additive SNP factor

An additional problem arises because the vector of weights γ is not observed and it is estimated in some sample. We define a polygenic index of y as follows:

$$\hat{g}_i = \frac{X_i \hat{\gamma}}{sd(X_i \hat{\gamma})} \tag{F.7}$$

In practice, $\hat{\gamma} \neq \gamma$ because the estimates of γ are based on final samples and the maximum level of predictive power is not achieved. We can write the standardized PGI as:

$$\hat{g}_i = \frac{g_i + \epsilon_i}{sd(g_i + \epsilon_i)} \tag{F.8}$$

, where $\epsilon_i = \frac{X_i(\hat{\gamma}-\gamma)}{sd(X_i\gamma)}$.

The predictive power of $\hat{g}_i$ is the R^2 of the regression of y_i on $\hat{g}_i$. Becker et al. (2021) show that $R^2 = \frac{h^2_{SNP}}{1+Var(e_i)} < h^2_{SNP}$. This suggests that when the weights are estimated with some error ($Var(\epsilon_i) \neq 0$), the predictive power of PGI is strictly lower than SNP heritability.

We use a similar notation as in Becker et al. (2021), and we define $\rho^2 = 1 + Var(e_i) = \frac{h^2_{SNP}}{R^2}$. Using this notation, PGI can be specified as $\hat{g}_i = \frac{g_i+\epsilon_i}{\rho}$.

If we estimate the effect of the additive SNP factor (g_i) on some outcome and instead of using g_i we use $\hat{g}_i$, the estimated coefficients will be biased. In what follows we characterize this bias in a model with own, siblings', and parental PGI, and we propose a bias correction methodology.

F.4 Bias Correction

Consider model (F.5) and suppose that we observe g_i, g_{si}, and g_{pi} with some error and the observed PGI are specified as:

$$\hat{g}_i = \frac{g_i + \epsilon_i}{\rho_g}; \qquad \hat{g}_{si} = \frac{g_{si} + \epsilon_{si}}{\rho_s}; \qquad \hat{g}_{pi} = \frac{g_{pi} + \epsilon_{pi}}{\rho_p} \tag{F.9}$$

, such that $Var(\epsilon_i) = \rho_g^2 - 1$, $Var(\epsilon_{si}) = \rho_s^2 - 1$, and $Var(\epsilon_{pi}) = \rho_p^2 - 1$. Also note that $Var(g_i) = Var(g_{si}) = Var(g_{pi}) = 1$, given that g is standardized.

Following Becker et al. (2021), we assume that ϵ is uncorrelated with all the other variables (g, Y, z). However, measurement error might be correlated among relatives. Moreover, the variance of the measurement error might be different in different samples.[11] In this section we assume that sibling and child-parent correlations of g and ρ are known. In Section F.7 we discuss how we set values for θ and ρ. Specifically, let us assume that the true sibling and child-parent genetic correlations are $Corr(g_i, g_{si}) = \theta$, $Corr(g_i, g_{pi}) = Corr(g_{si}, g_{pi}) = \theta_p$ (note that since the genetic variance is standardized to 1, the correlations are equal to the covariances). This implies that:

$$\begin{aligned} Cov(\epsilon_i, \epsilon_{si}) &= \rho_g \rho_s Cov(\hat{g}_i, \hat{g}_{si}) - \theta \\ Cov(\epsilon_i, \epsilon_{pi}) &= \rho_g \rho_p Cov(\hat{g}_i, \hat{g}_{pi}) - \theta_p \\ Cov(\epsilon_{si}, \epsilon_{pi}) &= \rho_s \rho_p Cov(\hat{g}_{si}, \hat{g}_{pi}) - \theta_p. \end{aligned} \tag{F.10}$$

We further denote the variance-covariance matrix of $(\epsilon_i, \epsilon_{si}, \epsilon_{pi})'$ by $Var(\mathbb{E}_i)$.

[11] This may happen, for instance, if genetic markers have different effects for different cohorts, and the GWAS in which the weights are estimated is based on a cohort different from the one in which the polygenic indexes are constructed.

We first derive the coefficients from the correct model defined by equation (F.5). Denote $\alpha_g = (\beta_g, \beta_s, \beta_p, \delta'_g)'$. OLS estimates of α_g are:

$$\hat{\alpha}_g = \begin{pmatrix} Var(G_i) & Cov(G_i, z_i) \\ Cov(G_i, z_i) & Var(z_i) \end{pmatrix}^{-1} \begin{pmatrix} Cov(G_i, Y_i) \\ Cov(z_i, Y_i) \end{pmatrix} = V_g^{-1} \begin{pmatrix} Cov(G_i, Y_i) \\ Cov(z_i, Y_i) \end{pmatrix} \tag{F.11}$$

, where $G_i = (g_i, g_{si}, g_{pi})'$ and $Var(G_i)$ is the variance-covariance matrix of G_i.

Now consider a model in which G_i is measured with error. The model can be written as:

$$Y_i = \beta_{\hat{g}} \hat{g}_i + \beta_{\hat{s}} \hat{g}_{si} + \beta_{\hat{p}} \hat{g}_{pi} + z_i \delta_{\hat{g}} + \nu_i \tag{F.12}$$

Define $\alpha_{\hat{g}} = (\beta_{\hat{g}}, \beta_{\hat{s}}, \beta_{\hat{p}}, \delta_{\hat{g}})'$ and $\hat{G}_i = (\hat{g}_i, \hat{g}_{si}, \hat{g}'_{pi})$. Then, OLS estimates of $\alpha_{\hat{g}}$ are:

$$\hat{\alpha}_{\hat{g}} = \begin{pmatrix} Var(\hat{G}_i) & Cov(\hat{G}_i, z_i) \\ Cov(\hat{G}_i, z_i) & Var(z_i) \end{pmatrix}^{-1} \begin{pmatrix} Cov(\hat{G}_i, Y_i) \\ Cov(z_i, Y_i) \end{pmatrix} = V_{\hat{g}}^{-1} P^{-1} \begin{pmatrix} Cov(G_i, Y_i) \\ Cov(z_i, Y_i) \end{pmatrix} \tag{F.13}$$

, where $P = \begin{pmatrix} diag(\rho) & 0_{|\mathbb{G} \times z|} \\ 0_{|z \times \mathbb{G}|} & I_{|z|} \end{pmatrix}$, $diag(\rho)$ is a diagonal matrix with ρ_g, ρ_s, ρ_p on the main diagonal, $I_{|z|}$ is an identity matrix of the size of z, and $0_{|z \times \mathbb{G}|}$ is a matrix of zeros of size $z \times \mathbb{G}$.

Note that (F.13) follows from:

$$\begin{pmatrix} Cov(\hat{G}_i, Y_i) \\ Cov(z_i, Y_i) \end{pmatrix} = P^{-1} \begin{pmatrix} Cov(G_i, Y_i) \\ Cov(z_i, Y_i) \end{pmatrix} \tag{F.14}$$

, where $\mathbb{E} = (\epsilon_i, \epsilon_{si}, \epsilon_{pi})'$.

Equation (F.13) implies that:

$$\hat{\alpha}_{\hat{g}} = V_{\hat{g}}^{-1} P^{-1} V_g \hat{\alpha}_g \tag{F.15}$$

V_g is unobserved, and therefore we want to define it in terms of moments of observables. Recall that $V_{\hat{g}} = \begin{pmatrix} Var(\hat{G}_i) & Cov(\hat{G}_i, z_i) \\ Cov(\hat{G}_i, z_i) & Var(z_i) \end{pmatrix}$ and $Var(\hat{G}_i) = \begin{pmatrix} Var(\hat{g}_i) & Cov(\hat{g}_i, \hat{g}_{si}) & Cov(\hat{g}_i, \hat{g}_{pi}) \\ & Var(\hat{g}_{si}) & Cov(\hat{g}_{si}, \hat{g}_{pi}) \\ & & Var(\hat{g}_{pi}) \end{pmatrix} =$ $P^{-1}\big(Var(G_i) + Var(\mathbb{E})\big)P^{-1}$.

Let us define matrix Σ as a partitioned matrix of the same size as V_g with the first block being $Var(\mathbb{E}_i)$, such that $\Sigma = \begin{pmatrix} Var(\mathbb{E}_i) & 0_{|\mathbb{E}\times z|} \\ 0_{|z\times\mathbb{E}|} & 0_{|z|} \end{pmatrix}$.

Therefore,

$$V_{\hat{g}} = P^{-1}(V_g + \Sigma)P^{-1} \tag{F.16}$$

and

$$V_g = PV_{\hat{g}}P - \Sigma \tag{F.17}$$

This implies that:

$$\hat{\alpha}_{\hat{g}} = V_{\hat{g}}^{-1}P^{-1}(PV_{\hat{g}}P - \Sigma)\hat{\alpha}_g = (P - V_{\hat{g}}^{-1}P^{-1}\Sigma)\hat{\alpha}_g \tag{F.18}$$

If $\hat{\alpha}_g$ is a consistent estimator of α_g, then $plim\ \hat{\alpha}_{\hat{g}} = (P - V_{\hat{g}}^{-1}P^{-1}\Sigma)\alpha_g$. Therefore, the corrected estimator of α_g can be computed as:

$$\alpha_g^{corr} = (P - V_{\hat{g}}^{-1}P^{-1}\Sigma)^{-1}\hat{\alpha}_{\hat{g}} = A\hat{\alpha}_{\hat{g}} \tag{F.19}$$

, where $A = (P - V_{\hat{g}}^{-1}P^{-1}\Sigma)^{-1}$.

Note that when: (i) there are no covariates, (ii) $\rho_g = \rho_s = \rho_p = \rho$, and (iii) the true genetic correlation is equal to the observed genetic correlation $\Big(Corr(\hat{g}_i, \hat{g}_{si}) = Corr(g_i, g_{si})$, $Corr(\hat{g}_i, \hat{g}_{pi}) = Corr(g_i, g_{pi})\Big)$, then $\alpha_g^{corr} = \rho\hat{\alpha}_{\hat{g}}$. This is similar to the "rule of thumb" derived in Becker et al. (2021).

F.5 Standard Errors

Equation F.19 implies that:

$$Var(\alpha_g^{corr}) = AVar(\hat{\alpha}_{\hat{g}})A' \tag{F.20}$$

Standard errors can be computed by taking the square root of the elements in the main diagonal of this matrix.

F.5.1 Bias in the Standard Errors

In this section we show that the corrected standard errors obtained in the regression with measurement error will generally differ from the standard errors estimated in the "true" model given by (F.5). In order to figure out what happens to the estimated standard errors first consider the following model:

$$Y_i = \hat{\beta}_{\hat{g}}\hat{g}_i + \hat{\beta}_{\hat{s}}\hat{g}_{si} + \hat{\beta}_{\hat{p}}\hat{g}_{pi} + \hat{\delta}_{\hat{g}} z_i + \hat{v}_i = \hat{\beta}_{\hat{g}}\frac{g_i + \epsilon_i}{\rho_g} + \hat{\beta}_{\hat{s}}\frac{g_{si} + \epsilon_{si}}{\rho_s} + \hat{\beta}_{\hat{p}}\frac{g_{pi} + \epsilon_{pi}}{\rho_p} + \hat{\delta}_{\hat{g}} z_i + \hat{v}_i \quad (F.21)$$

Now let us compute the residual:

$$\hat{v}_i = Y_i - \hat{\beta}_{\hat{g}}\frac{g_i + \epsilon_i}{\rho_g} - \hat{\beta}_{\hat{s}}\frac{g_{si} - \epsilon_{si}}{\rho_s} - \hat{\beta}_{\hat{p}}\frac{g_{pi} - \epsilon_{pi}}{\rho_p} - \hat{\delta}_{\hat{g}} z_i \quad (F.22)$$

Using equation (F.5) we obtain:

$$\begin{aligned}\hat{v}_i = \beta_g g_i + \beta_s g_{si} + \beta_p g_{pi} + z_i\delta_g + u_i - \hat{\beta}_{\hat{g}}\tfrac{g_i+\epsilon_i}{\rho_g} - \hat{\beta}_{\hat{s}}\tfrac{g_{si}-\epsilon_{si}}{\rho_s} - \hat{\beta}_{\hat{p}}\tfrac{g_{pi}-\epsilon_{pi}}{\rho_p} - \hat{\delta}_{\hat{g}} z_i = \\ g_i(\beta_g - \tfrac{\hat{\beta}_{\hat{g}}}{\rho_g}) + g_{si}(\beta_s - \tfrac{\hat{\beta}_{\hat{s}}}{\rho_s}) + g_{pi}(\beta_p - \tfrac{\hat{\beta}_{\hat{p}}}{\rho_p}) + z_i(\delta_{\hat{g}} - \hat{\delta}_g) + u_i - \tfrac{\hat{\beta}_{\hat{g}}}{\rho_g}\epsilon_i - \tfrac{\hat{\beta}_{\hat{s}}}{\rho_s}\epsilon_{si} - \tfrac{\hat{\beta}_{\hat{p}}}{\rho_p}\epsilon_{pi}\end{aligned} \quad (F.23)$$

Rewriting this in matrix form yields the following expression:

$$\hat{v}_i = (\alpha_g - P^{-1}\hat{\alpha}_{\hat{g}})' \begin{pmatrix} G_i \\ z_i \end{pmatrix} - \hat{\beta}' diag(\rho)^{-1}\mathbb{E} + u_i \quad (F.24)$$

, where $\hat{\beta} = (\hat{\beta}_g, \hat{\beta}_s, \hat{\beta}_p)'$.

The residual contains two additional sources of variation compared to the true error (u_i). The first term is due to the fact that α_g is biased towards zero. The second term is due to the additional variance introduced by the presence of measurement error in G_i. Since the vectors of random variables $(G_i, z_i)', \mathbb{E}$, and u_i are uncorrelated by assumption, it follows that:

$$plim\ Var(\hat{v}_i) = (\alpha_g - P^{-1}A^{-1}\alpha_g)'V_g(\alpha_g - P^{-1}A^{-1}\alpha_g) + (A^{-1}\alpha_g)'P^{-1}\Sigma P^{-1}A^{-1}\alpha_g + Var(u_i) \quad (F.25)$$

Under homoskedasticity, $plim\ nVar(\hat{\alpha}_g) = Var(u_i)V_g^{-1}$ and

$$\begin{aligned}&plim\ nVar(\alpha_g^{corr}) = plim\ nAVar(\hat{\alpha}_{\hat{g}})A' = plim\ AVar(\hat{v}_i)V_{\hat{g}}^{-1}A' = \\ &A\Big((\alpha_g - P^{-1}A^{-1}\alpha_g)'V_g(\alpha_g - P^{-1}A^{-1}\alpha_g) + (A^{-1}\alpha_g)'P^{-1}\Sigma P^{-1}A^{-1}\alpha_g + Var(u_i)\Big) \\ &\Big(P^{-1}(V_g + \Sigma)P^{-1}\Big)^{-1} A' \neq Var(u_i)V_g^{-1}\end{aligned} \quad (F.26)$$

, where n is the number of observations.

This implies that the variance of the corrected estimator $Var(\alpha^{corr})$ in general does not converge in probability to the variance of the coefficients estimated in the "true" model $Var(\hat{\alpha}_g)$.

By means of simulations, we demonstrate that the corrected t-statistics will be biased towards zero.

F.6 Monte-Carlo Simulations

In order to test how well the error correction procedure proposed in the previous sections adjusts the estimated coefficients and standard errors, we conduct a battery of Monte-Carlo simulations.

We simulate data for families. Specifically, we assume that some families may have two pairs of siblings (pair of siblings 1 and 2, and pair of siblings 2 and 3) and therefore they would appear in our sample twice. We generate the data such that 90% of families have one pair of siblings and 10 % have two pairs.

Our data generation process assumes that the outcome y depends on genetic variables G_1 (own PGI), G_2 (siblings' PGI), and G_p (parental PGI), as well as on control variables z_1 and z_2. Both z_1 and z_2 are simulated as standard normal random variables. We assume that G_1, G_2, and G_p are random variables with mean zero and variance one. The covariance between G_1 and G_2 is imposed to be equal to θ set at 0.56 (the theoretical genetic correlation between siblings assumed in our analysis given the share of MZ twins). The covariance between G_i and G_p is imposed to be $\theta_p = 1/\sqrt{2}$ for $i = 1, 2$ (the child-parent correlation in PGI defined in Appendix A4 of Trejo and Domingue 2018). Given that the sibling genetic correlation is θ, the intrafamily correlations of G_1 and G_2 are equal to θ, and the intrafamily correlation of G_p is 1.

The outcome y is generated as follows:

$$y_i = \gamma_1 G_{1i} + \gamma_2 G_{2i} + \gamma_p G_{pi} + \alpha_1 z_{1i} + \alpha_2 z_{2i} + \omega_i, \text{ where } \omega_i \sim N(0, 1)$$

We simulate y for several values of $\gamma_1, \gamma_2, \gamma_p$. The baseline values are imposed to be $\gamma_1 = 0.05$, $\gamma_2 = 0.2$, $\gamma_p = 0$, and $\alpha_1 = \alpha_2 = 0.2$. We use these as the baseline values because they are similar to the effects of PGI that we estimate in our data. In this baseline scenario, the obtained y_i has mean zero and standard deviation 1.06. We then standardize y so that it has mean zero and variance one. Therefore, in the correct standardized regressions, the estimated effects are approximately equal to the assumed effects divided by the standard deviation of y. For comparison, we also conduct simulations for combinations of $\gamma_1 = (-0.2, 0.05, 0.2)$, $\gamma_2 = (-0.2, 0.2)$, and $\gamma_p = (-0.2, 0.2)$. We also assume that the error term ω_i is correlated across family members and that its intraclass correlation is 0.5.

We generate $\hat{G}_1$, $\hat{G}_2$, and $\hat{G}_p$ as follows:

$$\hat{G}_{1i} = \frac{G_{1i} + \epsilon_{1i}}{\rho}, \ \hat{G}_{2i} = \frac{G_{2i} + \epsilon_{2i}}{\rho} \text{ and } \hat{G}_{pi} = \frac{G_{pi} + \epsilon_{pi}}{\rho}$$

, where ϵ_1, ϵ_2, ϵ_p are generated as correlated random variables with mean zero, variance $\rho^2 - 1$, $Corr(\epsilon_1, \epsilon_2) = \rho^2 Cov(\hat{G}_1, \hat{G}_2) - \theta$, and $Corr(\epsilon_i, \epsilon_p) = \rho^2 Cov(\hat{G}_i, \hat{G}_p) - \theta_p$ as derived in equation (F.10). We set $\rho = 1.4$, which is similar to the value of ρ for EA PGI estimated in Becker et al. (2021) using the Health and Retirement Study (HRS). We also conduct simulations for $\rho = 1.6$.

Next, we regress y on $\hat{G}_1$, $\hat{G}_2$, $\hat{G}_p$ z_1, z_2 and compute the corrected estimated effects, standard errors, and $t-$statistics.

Simulation results for different values of ρ based on 500 simulated samples are shown in Figure F.1. Simulation results for different effect sizes of G_1 assuming that the effect of G_2 is positive and the effect of G_p is zero are depicted in Figure F.2. Simulation results for different effect sizes of G_p assuming that the effects of G_1 and G_2 are positive are depicted in Figure F.3.

The simulation results obtained indicate that:

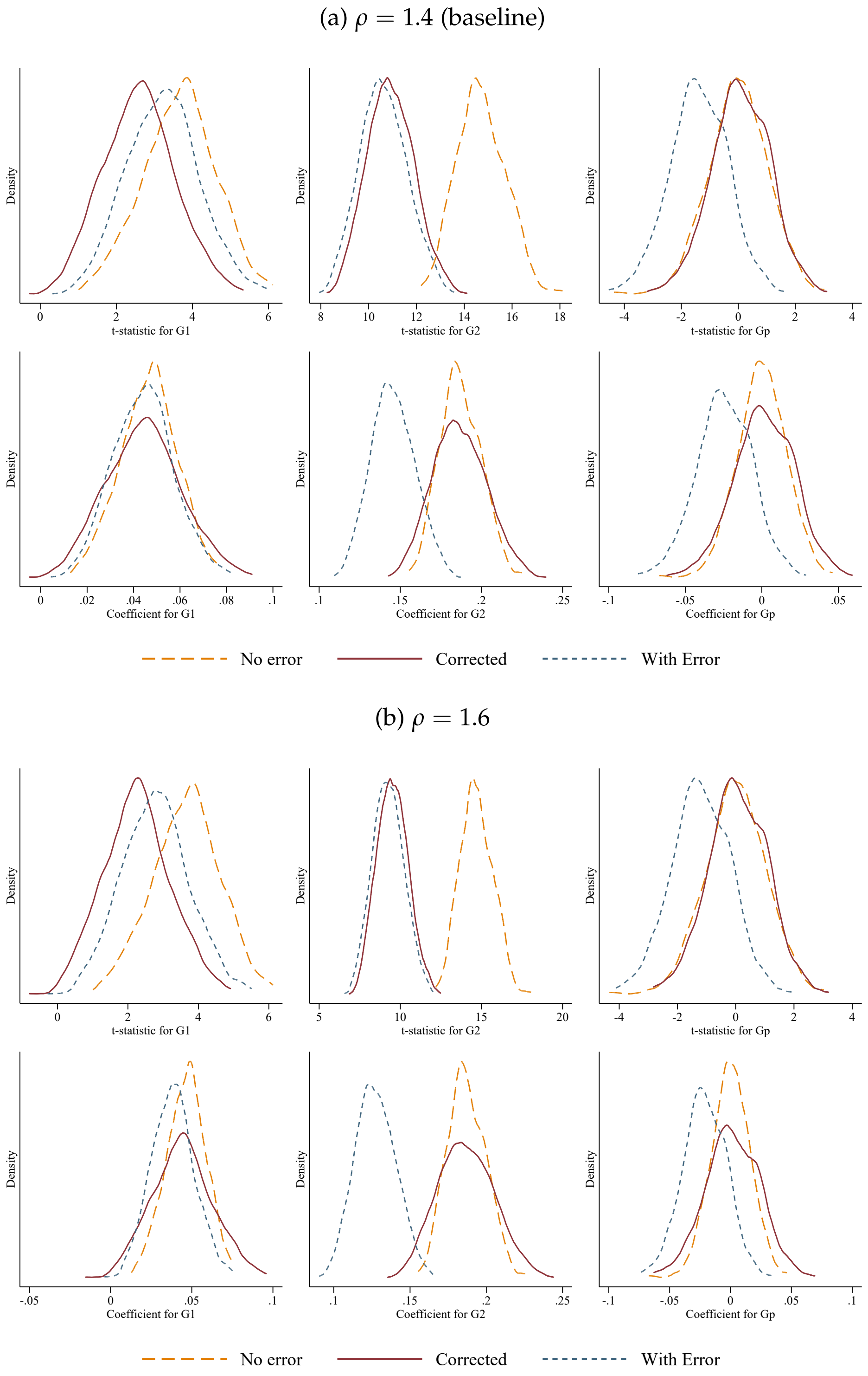
1. The uncorrected OLS regressions ("with error") yield substantial attenuation bias in both regression coefficients and $t-$statistics when the effect size is large.

2. The corrected coefficients on average match the correct regression results ("no error"). The corrected $t-$statistics are biased towards zero with respect to the regression with no errors.

3. Panel (a) of Figure F.1 shows that the attenuation bias in the estimated coefficients increases with ρ.

4. Similarly, Figures F.2 and F.3 show that the correction method works well for the different effect sizes considered.

Figure F.1: Simulation Results by Value of ρ. Kernel Densities of the Effect Sizes and t-statistics.

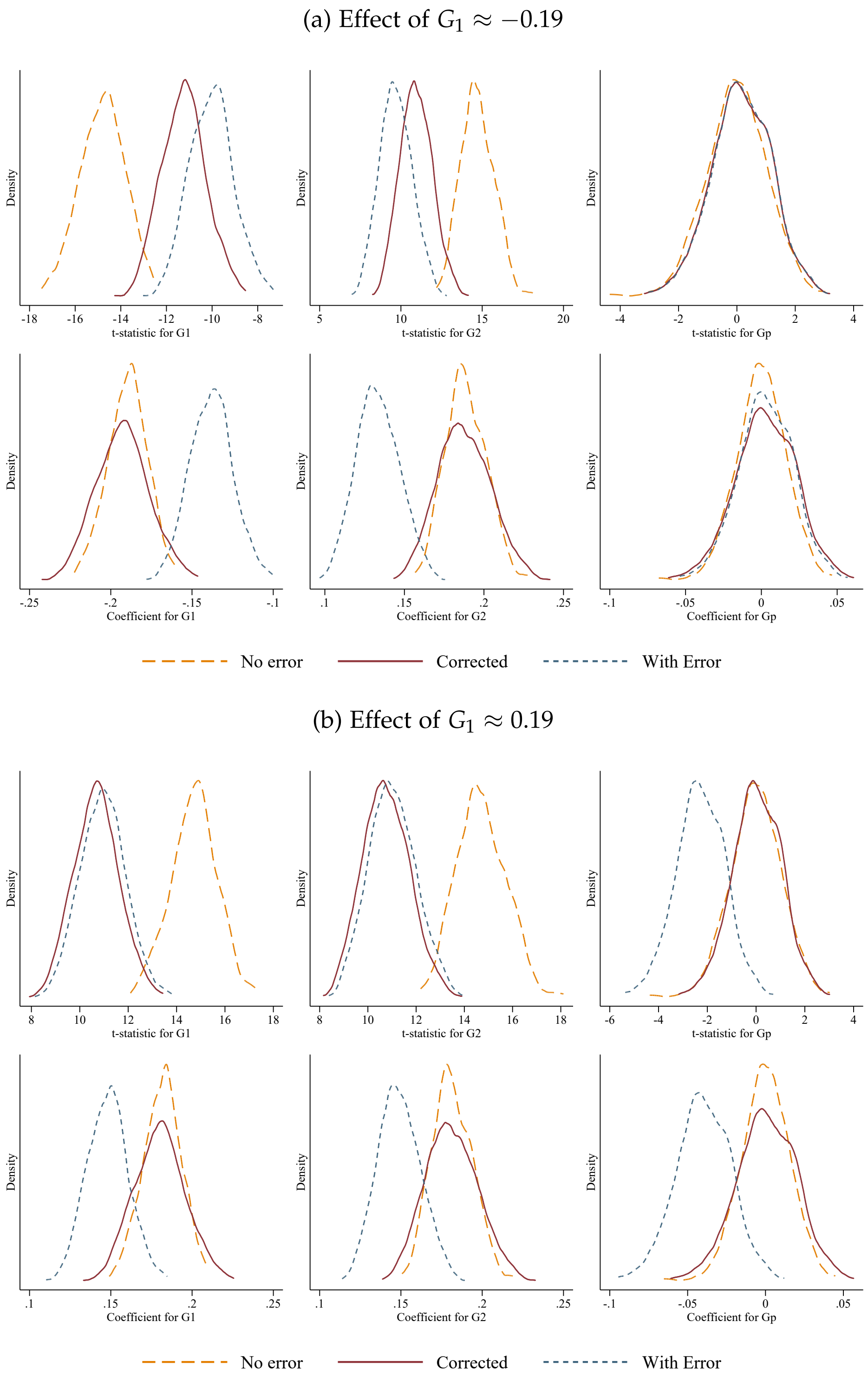


(a) $\rho = 1.4$ (baseline)

(b) $\rho = 1.6$

Note: This figure displays kernel-smoothed densities of the coefficients for G_1, G_2, and G_p obtained in 500 simulated samples of 10,000 families. 90% of families have one siblings pair and 10% of families have two siblings pairs. The results show the estimates obtained in the regressions with no measurement error, with measurement error, and when correcting for measurement error. The effect of $G_1 \approx 0.05$, the effect of $G_2 \approx 0.19$, and the effect of G_p is zero.

Figure F.2: Simulation Results by the Effect of G_1 for a Positive Effect of G_2, and a Zero Effect of G_p. Kernel Densities of the Effect Sizes and t-statistics.

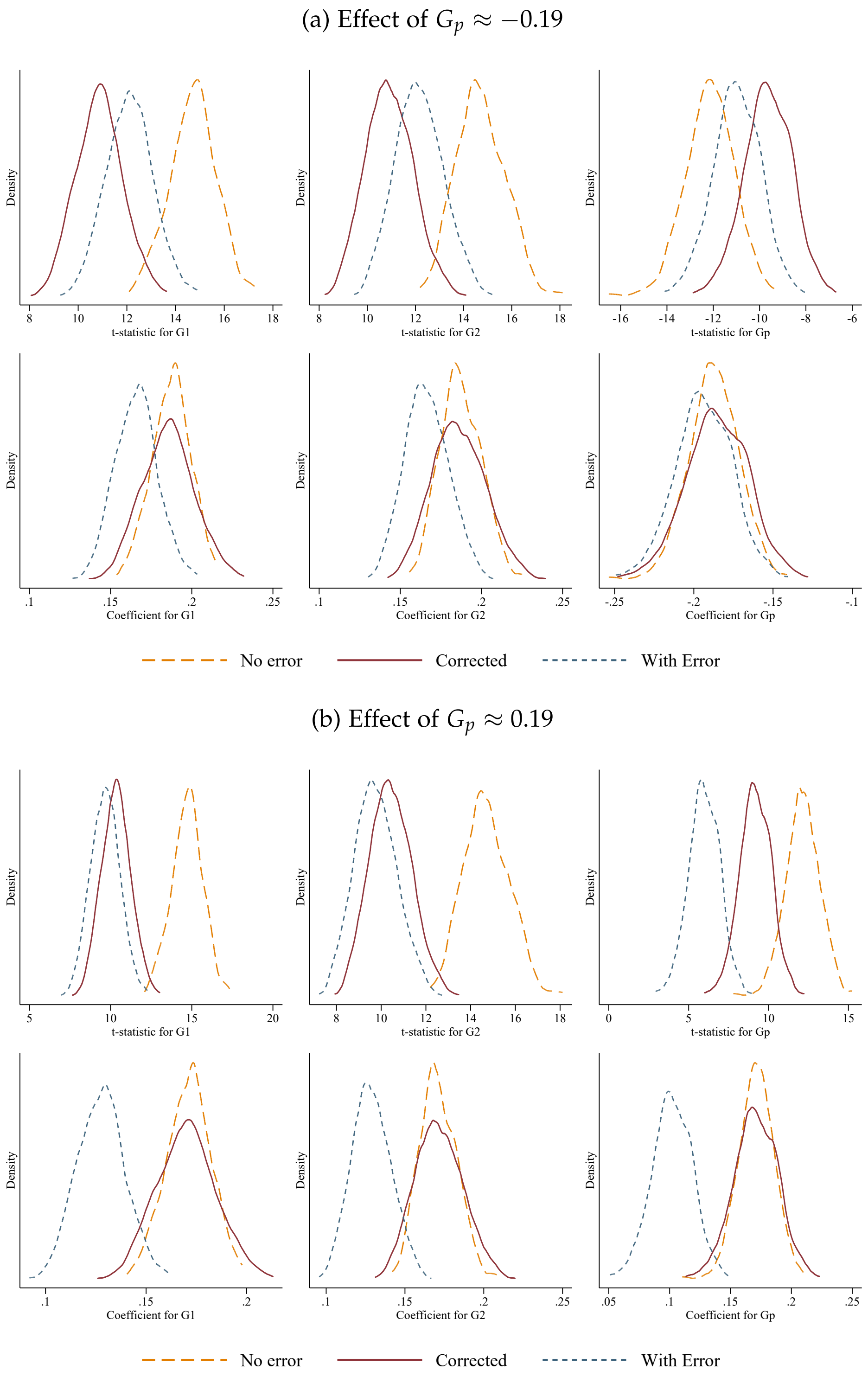


(a) Effect of $G_1 \approx -0.19$

(b) Effect of $G_1 \approx 0.19$

Note: This figure displays kernel-smoothed densities of coefficients for G_1, G_2, and G_p obtained in 500 simulated samples of $10,000$ families. 90% of families have one sibling pair and 10% of families have two sibling pairs. The results show the estimates obtained in the regressions with no measurement error, with measurement error, and when correcting for measurement error. The effect of $G_2 \approx 0.19$, the effect of G_p is zero, and $\rho = 1.4$.

Figure F.3: Simulation Results by the Effect of G_p for Positive Effects of G_1 and G_2. Kernel Densities of the Effect Sizes and t-statistics.



(a) Effect of $G_p \approx -0.19$

(b) Effect of $G_p \approx 0.19$

Note: This figure displays kernel-smoothed densities of coefficients for G_1, G_2, and G_p obtained in 500 simulated samples of 10,000 families. 90% of families have one sibling pair and 10% of families have two sibling pairs. The results show the estimates obtained in the regressions with no measurement error, with measurement error, and when correcting for measurement error. Effects of G_1 and $G_2 \approx 0.19$; $\rho = 1.4$.

F.7 Choice of Parameters

The measurement error method previously described requires an assumption about the SNP heritability of educational attainment (h^2_{SNP}), which is used to compute ρ_g, ρ_s, and ρ_p, and about the sibling and parent-child correlations of the additive SNP factors ($\theta = Corr(g_i, g_{si})$ and $\theta_p = Corr(g_i, g_{pi})$.

To infer θ, we estimate pair-wise kinship coefficients in our sample of siblings, excluding monozygotic (MZ) twins. We use the algorithm proposed by Manichaikul et al. (2010) for robust relationship inference in genome-wide association studies.[12] The estimated average pair-wise kinship coefficient is equal to 0.249, which implies that the sibling genetic correlation is equal to 0.498. Given that the full-sibling genetic correlation is not statistically distinguishable from 0.5, we impose the assumption that there is no assortative mating, which implies that the parents-child genetic correlation is $\theta_p = \frac{1}{\sqrt{2}}$ (Trejo and Domingue, 2018). We compute $\theta = 0.5 \times (Share\ non - MZ) + 1 \times (Share\ MZ)$, given that MZ twins have perfect genetic correlation.

We estimate h^2_{SNP} for educational attainment and cognitive performance using the genome-wide complex trait analysis tool (GCTA) (Yang et al., 2011). GCTA estimates the variance of the trait explained by all measured SNPs instead of testing the association between any particular SNP and the trait. We apply GCTA to the Add Health sample of European ancestry individuals with available genetic data. Yang et al. (2011) recommend excluding close relatives from the analysis, since common environmental effects can inflate the estimates of the genetic variance. Following their recommendation, we exclude individuals with genetic relatedness (estimated from genome-wide data) greater than 0.025. This leaves us with 4,818 unrelated individuals used for the estimation of h^2_{SNP}. One concern with the Add Health dataset is that it is a school-based survey, so individuals are more likely to share a common environment. If there is some genetic sorting into schools, then GCTA estimates of the variance of educational attainment and cognitive performance explained by the SNPs may be inflated.[13] In order to avoid this issue, we remove school fixed effects from educational attainment and cognitive performance and use the obtained residuals as phenotypes for GCTA estimation of h^2_{SNP}.

For educational attainment, the obtained estimate of h^2_{SNP} is equal to 0.229 ($SE = 0.070$) when school fixed effects are adjusted for. In contrast, the estimate of h^2_{SNP} unadjusted for school selection amounts to 0.471 ($SE = 0.069$), which is significantly larger than the estimates

[12] The algorithm is implemented in a publicly available software package, KING, downloadable from `https://www.kingrelatedness.com/`.

[13] Domingue et al. (2018) document genetic similarity among friends that is mainly driven by non-random school assignment in the Add Health sample. Yengo et al. (2020) suggest that these results are mainly driven by uncontrolled population stratification.

of h^2_{SNP} obtained in other data sets provided in Supplementary Table 4 of Becker et al. (2021). Therefore, we use 0.229 as a benchmark estimate of h^2_{SNP} for educational attainment. The estimated R^2 in the regression of educational attainment on the EA PGI net of school fixed effects is equal to 0.080, which implies an estimated baseline value of $\rho \approx \sqrt{0.229/0.080} \approx 1.693$. The value of ρ unadjusted for school selection is $\rho \approx \sqrt{0.471/0.122} \approx 1.968$ (the R^2 of the regression of educational attainment on the EA PGI is 0.122).

We use the Peabody Picture Vocabulary Test (PPTV) score to measure cognitive ability. For PPVT, the estimated $h^2_{SNP} = 0.156$ in the analysis adjusted for school fixed affects and to $h^2_{SNP} = 0.341$ in unadjusted analysis. $R^2 = 0.0471$ and $R^2 = 0.0549$ when school fixed effects are adjusted and when they are unadjusted respectively. Therefore, for cognitive performance, the baseline value of ρ is 1.819 and the unadjusted for school effects value is $\rho = 2.492$.

Becker et al. (2021) advise users to rely on an estimate of ρ from a larger dataset if their sample size is too small. They provide estimates of ρ from three large samples: the Health and Retirement Study (HRS, $N = 12,090$), the Wisconsin Longitudinal Study (WLS,$N = 8,949$), and the UK Biobank-3rd partition (UKB, $N = 145,960$). Their estimates of ρ for educational attainment in the HRS, the WLS and the UKB amount to 1.413, 1.649, and 1.452, respectively. Their estimates of ρ for cognitive performance in WLS and the UKB are 1.991 and 1.697 respectively. We therefore use these values in sensitivity checks provided in Table H.10 of the Appendix.

G Socioeconomic Index Construction

We use information on parental education, parental occupation, household income, household receipt of public assistance, and residential building quality to construct a family socioeconomic status (SES) index.

We construct parental educational attainment using the question *"How far did you go in school?"* addressed to parents in Wave I, as well as the question *"How far in school did she(he) [mother (father)] go?"* addressed to children in Wave I about their parents. Maternal/paternal educational attainment is based on parents' own answers if they participated in the parental interview, and on their child's answers otherwise. Parental educational attainment is defined as the average of paternal and maternal educational attainment.

We use children's answers to the question *"What kind of work does she (he) [mother (father)] do?"* regarding parental occupation. We assign occupational prestige scores based on the National Opinion Research Center (NORC) occupational classification.[14] We then compute the

[14] http://ibgwww.colorado.edu/~agross/NNSD/prestige%20scores.html

parental occupational prestige score as the average of mother's and father's prestige scores. If both parents have no occupation, the value of the index is set to zero.

Family income is based on the following question addressed to parents in Wave I: *"About how much total income, before taxes did your family receive in 1994? Include your own income, the income of everyone else in your household, and income from welfare benefits, dividends, and all other sources."*. If family income is missing, we impute missing values using information on residential building quality, gender, ethnicity, race, and parental education. We then take the *log* of family income (adding 1 in order to avoid missing values if family income is zero).

As for household receipt of public assistance, we rely on the following question asked to children in Wave I: *"Does she (he) [mother (father)] receive public assistance, such as welfare?"*. We then compute a parent on welfare indicator as the average of mother's and father's indicators.

We construct a residential building quality variable using the question *"How well kept is the building in which the respondent lives?"* reported by interviewers. We create a dummy variable "good quality residential building", which takes the value one if the answer was "very well kept" or "fairly well kept", and zero if the answer was "poorly kept" or "very poorly kept".

Finally, we conduct a principal component analysis of parental education, parental occupational prestige, family income, household receipt of public assistance, and residential building quality to produce a obtain index. The first principal component explains 44.8% of the variance of these variables. We use loadings on this component to compute a SES index, and then we standardize it to have mean 0 and standard deviation 1.

H Tables and Figures

Table H.1: Summary Statistics of Regressors

	Mean	Std. Dev
Polygenic Indexes		
EA PGI	0.000	1.000
Sibling's EA PGI	0.000	1.000
Parental EA PGI	0.000	1.000
EA PGI-Sibling's EPGI	0.000	0.936
CP PGI	0.000	1.000
Sibling's CP PGI	0.000	1.000
Parental CP PGI	0.000	1.000
CP PGI-Sibling's CP PGI	0.000	0.987
Baseline Controls		
Age	16.770	1.477
Age squared	283.419	48.753
Age-Sibling's age (months)	19.290	17.025
Female	0.508	0.500
Female sibling	0.533	0.499
Additional Controls		
Rural	0.342	0.473
Both parents live in household	0.776	0.417
SES index	0.000	1.000
N	604	

Note: EA PGI is the educational attainment polygenic index, CP PGI is the cognitive performance polygenic index, and SES index is a socio-economic index constructed as described in Appendix G. PGI is always standardized to have mean 0 and standard deviation 1.

Table H.2: Parental Investment Indicators

	Mean	Std. Dev	N
Parental Investment Index	0.000	1.000	604
Maternal Investment Index	0.000	1.000	579
Paternal Investment Index	0.000	1.000	481
Parental Investment Index Componponents			
Maternal Investment Index Components			
Visited some event with mother. W1	0.271	0.445	579
Talk about school with mother. W1	0.644	0.479	579
Worked on a project with mother. W1	0.128	0.334	579
Talk about other school things with mother. W1	0.561	0.497	579
Paternal Investment Index Components			
Visited some event with father. W1	0.243	0.429	481
Talk about school with father. W1	0.543	0.499	481
Worked on a project with father. W1	0.081	0.273	481
Talk about other school things with father. W1	0.497	0.501	481
Additional Parental Index Components			
Nr days at least 1 parent present when eating evening meal in past 7d	4.456	2.458	603

Note: Parental investment indexes are standardized to have mean 0 and standard deviation 1.

Table H.3: Across- and Within-Family Standard Deviations of Parental Investment Indexes

	(1)	(2)	(3)	(4)	(5)	(6)
	Pooled Sample		Non-Twins		Twins	
	Across	Within	Across	Within	Across	Within
Parental Investment Index	1.000	0.587	1.005	0.620	0.990	0.509
Maternal Investment Index	0.998	0.579	0.999	0.616	0.997	0.487
Paternal Investment Index	1.001	0.582	0.998	0.606	1.007	0.527

Note: This table reports standard deviations of the parental investment indexes (columns 1, 3, and 5), and the same standard deviations when family fixed effects are removed (columns 2, 4, and 6).

Table H.4: Parental Investment Indexes and Educational Attainment in Wave IV

	(1)	(2)	(3)	(4)	(5)	(6)
Parental Investment Index	0.175	0.112				
	(0.039)	(0.034)				
	[<0.001]	[<0.001]				
Maternal Investment Index			0.167	0.106		
			(0.040)	(0.035)		
			[<0.001]	[0.003]		
Paternal Investment Index					0.187	0.133
					(0.043)	(0.038)
					[<0.001]	[0.001]
Controls	No	Yes	No	Yes	No	Yes
N	604	604	579	579	481	481
R^2	0.031	0.359	0.028	0.353	0.034	0.348

Note: This table reports OLS coefficient estimates of the association between educational attainment (standardized to have mean 0 and standard deviation 1) and parental investments (as measured by an index standardized to have mean 0 and standard deviation 1). The regressions include age, age-squared, a female dummy, a rural area dummy, an indicator that both parents cohabit, the SES index, and parental EA PGI. Standard errors clustered at the family level are in parentheses.

Table H.5: Summary Statistics for Baseline Characteristics in Estimation and Representative Samples

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Representative Sample		Only White		Estimation Sample		P-value for the difference	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	(1)-(5)	(3)-(5)
EA PGI	-	-	0.007	1.008	0.000	1.000	-	0.866
CP PGI	-	-	0.056	0.933	0.000	1.000	-	0.199
Black	0.164	0.370	0.000	-	0.000	-	-	-
Age	15.918	1.784	15.862	1.755	16.770	1.477	<0.001	<0.001
Female	0.490	0.500	0.489	0.500	0.508	0.500	0.380	0.373
Rural	0.281	0.447	0.333	0.469	0.342	0.473	0.002	0.665
Both parents live in household	0.710	0.454	0.775	0.418	0.776	0.417	<0.001	0.940
SES index	-0.028	0.982	-0.040	0.983	0.000	1.000	0.510	0.345
Parental Investment Index	-0.013	1.039	0.041	1.043	0.000	1.000	0.755	0.341
Maternal Investment Index	-0.017	1.007	0.040	1.011	0.000	1.000	0.696	0.360
Paternal Investment Index	0.050	1.059	0.103	1.060	0.000	1.000	0.290	0.033
N	18,523		9,452		604			

Note: EA PGI is the educational attainment polygenic index, CP PGI is the cognitive performance polygenic index, SES index is a socio-economic index constructed as described in Appendix G. The sample of "only white" individuals is obtained by using genetic race for individuals with available genetic information and self-reported race for those without available genetic information. EA PGI, CP PGI, SES index, and parental investment indexes are standardized using means and standard deviations of these variables in the estimation sample. Statistics from columns 1-4 are computed using the Add Health Wave I grand sample weights.

Table H.6: Balancing Tests

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Baseline Controls					Additional Controls		
	Age	Age squared	Age difference	Female	Female sibling	Rural	Both parents cohabit	SES
EA PGI-Sibling's EA PGI	0.022	0.016	0.090	0.036	0.022	-0.039	-0.077	-0.101
SE	(0.080)	(0.080)	(0.084)	(0.082)	(0.091)	(0.085)	(0.091)	(0.078)
$p - value$	[0.786]	[0.840]	[0.283]	[0.659]	[0.808]	[0.651]	[0.401]	[0.193]
EA PGI	0.002	0.003	-0.097	-0.057	-0.021	-0.012	0.114	0.271
SE	(0.081)	(0.080)	(0.076)	(0.078)	(0.080)	(0.075)	(0.080)	(0.072)
$p - value$	[0.976]	[0.967]	[0.200]	[0.472]	[0.793]	[0.877]	[0.157]	[<0.001]
Parental EA PGI	0.032	0.030	0.060	0.034	-0.110	-0.168	0.040	0.394
SE	(0.074)	(0.074)	(0.062)	(0.068)	(0.068)	(0.081)	(0.092)	(0.066)
$p - value$	[0.669]	[0.682]	[0.330]	[0.623]	[0.107]	[0.039]	[0.663]	[<0.001]
N	604	604	604	604	604	604	604	604

Note: EA PGI is the educational attainment polygenic index. This table displays OLS coefficients obtained after regressing each control variable on sibling differences in EA PGI, own EA PGI, and parental EA PGI. EA PGI is always standardized to have mean 0 and standard deviation 1. All individual and family characteristics are measured in Wave I of Add Health. Coefficient estimates and standard errors are measurement-error-corrected as described in Appendix F. Standard errors clustered at the family level are in parentheses.

Table H.7: The Effect of Educational Polygenic Index and Sibling Differences in Educational Polygenic Indexes on Parental Investments. Non-corrected Results

	(1) Pooled Sample	(2) Non-Twins	(3) Twins
EA PGI-Siblings EA PGI	-0.222	-0.213	-0.123
SE	(0.084)	(0.109)	(0.158)
$p - value$	[0.009]	[0.052]	[0.439]
EA PGI	0.294	0.227	0.361
SE	(0.150)	(0.203)	(0.223)
$p - value$	[0.051]	[0.264]	[0.108]
Parental EA PGI	-0.159	-0.100	-0.233
SE	(0.136)	(0.174)	(0.235)
$p - value$	[0.243]	[0.564]	[0.322]
N	604	412	192

Note: EA PGI is the educational attainment polygenic index. This table reports OLS estimated effects of sibling difference in EA PGI, own EA PGI, and parental EA PGI on parental investments (as measured by an index standardized to have mean 0 and standard deviation 1) in the pooled sample (1), in the sample of non-twins (2), and in the sample of twins (3). EA PGIs is always standardized to have mean 0 and standard deviation 1. The regressions include age, age-squared, sibling differences in age (only included in the non-twin sample), a female dummy, and a female sibling dummy. Standard errors clustered at the family level are in parentheses.

Table H.8: Heterogeneous Effects of Educational Polygenic Index and Sibling Differences in Educational Polygenic Indexes on Parental Investments

	(1)	(2)	(3)	(4)
	Maternal Investment	Paternal Investment	Only Females	Only Males
EA PGI-Sibling's EA PGI	-0.220	-0.163	-0.249	-0.220
SE	(0.083)	(0.096)	(0.113)	(0.118)
$p - value$	[0.008]	[0.090]	[0.028]	[0.064]
EA PGI	0.177	0.179	0.300	0.150
SE	(0.081)	(0.080)	(0.106)	(0.118)
$p - value$	[0.028]	[0.026]	[0.005]	[0.206]
Parental EA PGI	0.033	0.019	-0.002	0.013
SE	(0.080)	(0.072)	(0.104)	(0.102)
$p - value$	[0.677]	[0.792]	[0.988]	[0.895]
N	579	481	307	297

Note: EA PGI is the educational attainment polygenic index. This table reports the estimated effects of sibling difference in EA PGI, own EA PGI, and parental EA PGI on parental investments (as measured by an index standardized to have mean 0 and standard deviation 1) in the pooled sample. EA PGI is always standardized to have mean 0 and standard deviation 1. The regressions include age, age-squared, sibling differences in age (only included in the non-twin sample), a female dummy, and a female sibling dummy. Columns 1 and 2 report the results obtained when the parental investment index is based only on maternal and paternal investment variables respectively. The regressions in columns 3 and 4 report the estimated effects for female and male children separately. Coefficient estimates and standard errors are measurement-error-corrected as described in Appendix F. Standard errors clustered at the family level are in parentheses.

Table H.9: The Effect of Polygenic Indexes and Sibling Differences in Polygenic Indexes on Parental Investments. Sensitivity

	(1)	(2)	(3)	(4)	(5)	(6)
	Panel A: Cognitive Performance PGI			Panel B: MZ twins excluded		
	Pooled Sample	Non-Twins	Twins	Pooled Sample	Non-Twins	Twins
PGI-Sibling's PGI	-0.200	-0.235	-0.058	-0.229	-0.270	-0.068
SE	(0.092)	(0.116)	(0.171)	(0.082)	(0.100)	(0.137)
$p-value$	[0.031]	[0.044]	[0.735]	[0.006]	[0.007]	[0.622]
PGI	0.011	-0.078	0.130	0.204	0.174	0.303
SE	(0.072)	(0.090)	(0.118)	(0.085)	(0.095)	(0.184)
$p-value$	[0.878]	[0.384]	[0.270]	[0.017]	[0.067]	[0.103]
Parental PGI	0.046	0.087	0.059	-0.005	0.023	-0.064
SE	(0.075)	(0.096)	(0.111)	(0.081)	(0.093)	(0.155)
$p-value$	[0.540]	[0.367]	[0.596]	[0.946]	[0.800]	[0.679]
N	604	412	192	531	412	119
	Panel C: Controls for 20 Principal Components			Panel D: Extended Sample		
	Pooled Sample	Non-Twins	Twins	Pooled Sample	Non-Twins	Twins
PGI-Sibling's PGI	-0.244	-0.305	-0.014	-0.105	-0.128	0.002
SE	(0.087)	(0.106)	(0.160)	(0.054)	(0.067)	(0.088)
$p-value$	[0.005]	[0.004]	[0.932]	[0.051]	[0.055]	[0.981]
PGI	0.215	0.159	0.299	0.124	0.111	0.128
SE	(0.078)	(0.097)	(0.134)	(0.069)	(0.079)	(0.103)
$p-value$	[0.006]	[0.104]	[0.027]	[0.071]	[0.163]	[0.216]
Parental PGI	0.000	0.035	0.056	0.117	0.132	0.083
SE	(0.079)	(0.104)	(0.127)	(0.064)	(0.075)	(0.098)
$p-value$	[0.998]	[0.737]	[0.657]	[0.067]	[0.079]	[0.399]
N	604	412	192	1215	832	383

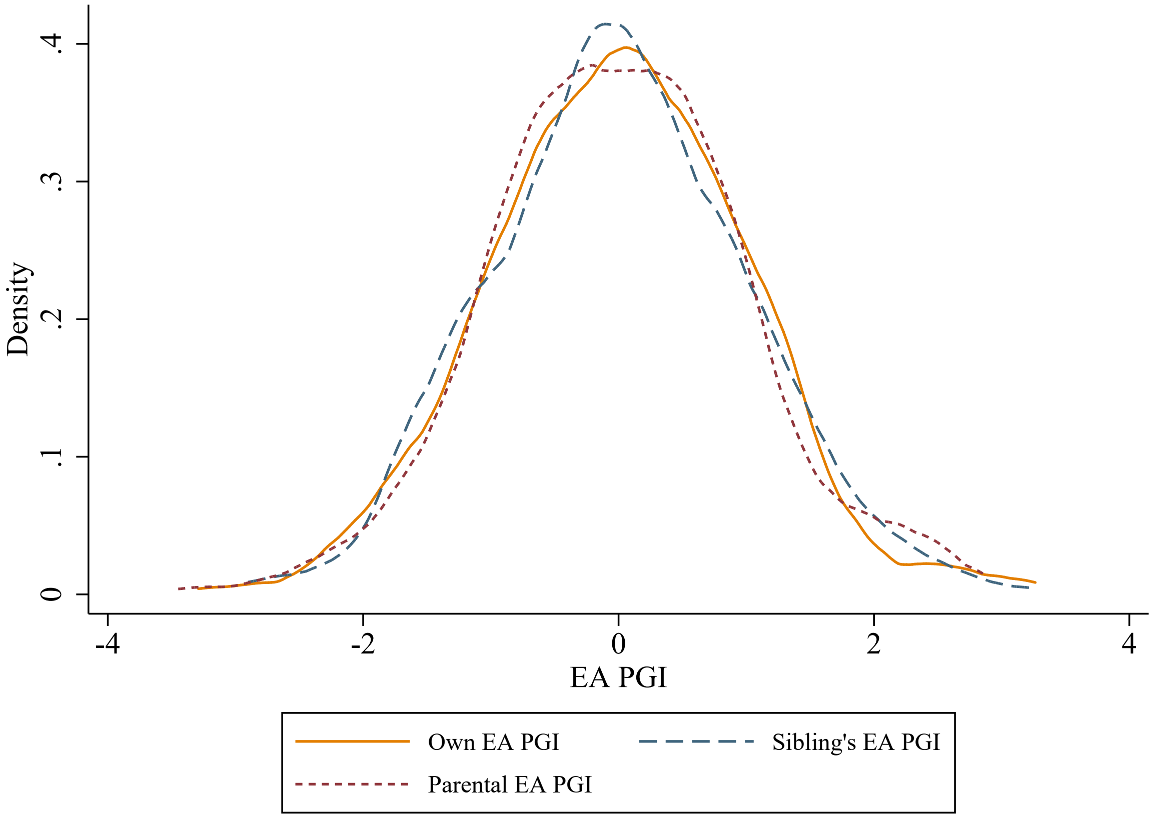
Note: PGI is the polygenic index. This table reports the estimated effects of sibling differences in PGI, own PGI, and parental PGI on parental investments (as measured by an index standardized to have mean 0 and standard deviation 1) in the pooled sample, in the sample of non-twins, and in the sample of twins. PGI is always standardized to have mean 0 and standard deviation 1. The regressions include age, age-squared, sibling differences in age (only included in the non-twin sample), a female dummy, and a female sibling dummy. Panel A uses cognitive performance polygenic index (CP PGI), while all other panels use educational attainment polygenic indexes (EA PGI). Panel B excludes monozygotic twins from the sample, Panel C includes the first 20 principal components of the full genetic relatedness matrix as controls, and Panel D expands the sample of our benchmark analysis (Table 3) to include firstborns and later-born siblings. Coefficient estimates and standard errors are measurement-error-corrected as described in Appendix F. Standard errors clustered at the family level are in parentheses.

Table H.10: The Effect of Polygenic Indexes and Sibling Differences in Polygenic Indexes on Parental Investments. Sensitivity to Different Values of Heritability

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	EA PGI with heritability estimated in:				CP PGI with heritability estimated in:		
	AH (no school FE)	HRS	WLS	UKB	AH (no school FE)	WLS	UKB
PGI-Sibling's PGI	-0.273	-0.194	-0.227	-0.200	-0.280	-0.220	-0.186
SE	(0.094)	(0.067)	(0.078)	(0.069)	(0.129)	(0.101)	(0.086)
$p - value$	[0.004]	[0.004]	[0.004]	[0.004]	[0.030]	[0.031]	[0.031]
PGI	0.247	0.177	0.206	0.181	0.015	0.012	0.010
SE	(0.090)	(0.065)	(0.075)	(0.066)	(0.100)	(0.079)	(0.067)
$p - value$	[0.006]	[0.006]	[0.006]	[0.006]	[0.882]	[0.879]	[0.877]
Parental PGI	0.000	-0.001	-0.001	-0.001	0.063	0.050	0.043
SE	(0.086)	(0.061)	(0.072)	(0.063)	(0.103)	(0.082)	(0.070)
$p - value$	[1.000]	[0.989]	[0.993]	[0.990]	[0.542]	[0.540]	[0.539]
N	604	604	604	604	604	604	604

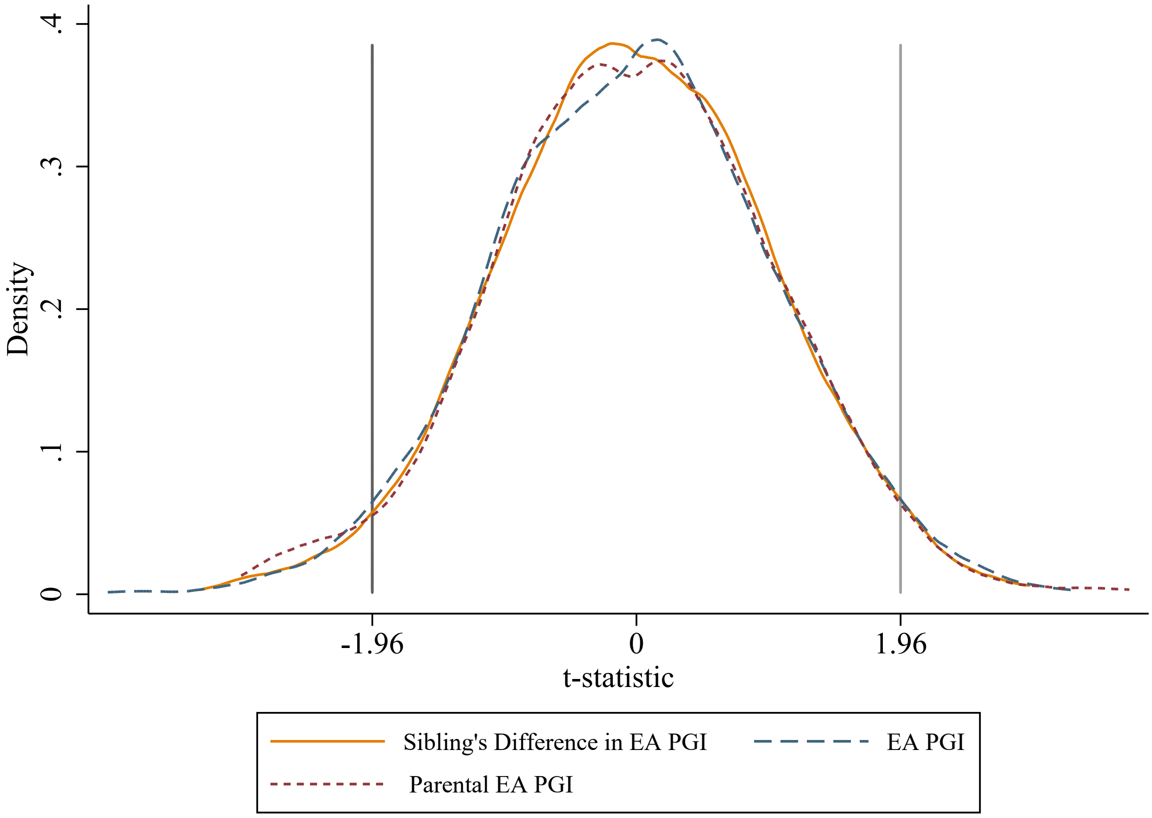
Note: EA PGI is the educational attainment polygenic index and CP PGI is the cognitive performance polygenic index. This table reports the estimated effects of sibling differences in PGI, own PGI, and parental PGI on parental investments (as measured by an index standardized to have mean 0 and standard deviation 1). PGI is always standardized to have mean 0 and standard deviation 1. The regressions include age, age-squared, sibling differences in age (only included in the non-twin sample), a female dummy, and a female sibling dummy. Coefficient estimates and standard errors are measurement-error-corrected as described in Appendix F. Columns 1-4 report the estimated effects of EA PGI and columns 5-7 report the effects of CP PGI using different values of R^2 and SNP heritability based on different data sources for the measurement-error correction. Columns 1 and 5 use estimates of SNP heritability for EA PGI and CP PGI based on the sample of unrelated individuals from Add Health (not adjusted for school fixed effects). The rest of the columns correct for measurement error using SNP heritability and R^2 estimates from Supplementary Table 4 of Becker et al. (2021) for the Health and Retirement Study (HRS), the Wisconsin Longitudinal Study (WLS), and the UK Biobank - 3rd partition (UKB3). Standard errors clustered at the family level are in parentheses.

Figure H.1: Educational Polygenic Indexes (Standardized). Kernel Density Estimates



Note: EA PGI is the educational attainment polygenic index. This figure displays kernel-smoothed densities of own EA PGI, siblings' EA PGI, and imputed parental EA PGI standardized to have mean 0 and standard deviation 1. No. observations: 604.

Figure H.2: Distribution of Placebo t-values



Note: EA PGI is the educational attainment polygenic index. This figure shows the distributions of the t-values of the tests $\beta_1 = 0$, $\beta_2 = 0$, and $\beta_3 = 0$ obtained when estimating 1000 placebo regressions of the parental investment index, where actual values of EA PGI, siblings' EA PGI, and parental EA PGI are replaced with those from randomly chosen families from our sample. The baseline controls listed in Table H.6 are included in the regressions. Coefficient estimates and standard errors are measurement-error-corrected as described in Appendix F. Standard errors are clustered at the family level.

References

Abdellaoui, A. and Verweij, K. J. (2021): "Dissecting polygenic signals from genome-wide association studies on human behaviour," *Nature Human Behaviour*, 5, 686–694.

Becker, J., Burik, C. A., Goldman, G. et al. (2021): "Resource profile and user guide of the Polygenic Index Repository," *Nature Human Behaviour*, 1–15.

Benjamin, D., Cesarini, D., Okbay, A. et al. (2021): "Polygenic Index Inventories Documentation," Tech. rep.

Chatterjee, N., Wheeler, B., Sampson, J. et al. (2013): "Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies," *Nature Genetics*, 45, 400–405.

Cheesman, R., Selzam, S., Ronald, A. et al. (2017): "Childhood behaviour problems show the greatest gap between DNA-based and twin heritability," *Translational Psychiatry*, 7, 1–9.

Demange, P. A., Malanchini, M., Mallard, T. T. et al. (2021): "Investigating the genetic architecture of noncognitive skills using GWAS-by-subtraction," *Nature Genetics*, 53, 35–44.

Domingue, B. W., Belsky, D. W., Fletcher, J. M. et al. (2018): "The social genome of friends and schoolmates in the National Longitudinal Study of Adolescent to Adult Health," *Proceedings of the National Academy of Sciences*, 115, 702–707.

Elliott, M. L., Belsky, D. W., Anderson, K. et al. (2019): "A polygenic score for higher educational attainment is associated with larger brains," *Cerebral Cortex*, 29, 3496–3504.

Falconer, D. and Mackay, T. (1996): "Introduction to quantitative genetics. Essex," *UK: Longman Group*.

Hill, W. G., Goddard, M. E. and Visscher, P. M. (2008): "Data and theory point to mainly additive genetic variance for complex traits," *PLoS Genetics*, 4, e1000008.

Hivert, V., Sidorenko, J., Rohart, F. et al. (2021): "Estimation of non-additive genetic variance in human complex traits from a large sample of unrelated individuals," *The American Journal of Human Genetics*, 108, 786–798.

International HapMap 3 Consortium and others (2010): "Integrating common and rare genetic variation in diverse human populations," *Nature*, 467, 52.

LEE, J. J., WEDOW, R., OKBAY, A. ET AL. (2018): "Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals," *Nature Genetics*, 50, 1112–1121.

LI, Y., WILLER, C., SANNA, S. ET AL. (2009): "Genotype imputation," *Annual Review of Genomics and Human Genetics*, 10, 387–406.

MANICHAIKUL, A., MYCHALECKYJ, J. C., RICH, S. S. ET AL. (2010): "Robust relationship inference in genome-wide association studies," *Bioinformatics*, 26, 2867–2873.

OKBAY, A., WU, Y., WANG, N. ET AL. (2022): "Polygenic prediction of educational attainment within and between families from genome-wide association analyses in 3 million individuals," *Nature Genetics*, 54, 437–449.

PALMER, D. S., ZHOU, W., ABBOTT, L. ET AL. (2021): "Analysis of genetic dominance in the UK Biobank," *bioRxiv*.

RIETVELD, C. A., MEDLAND, S. E., DERRINGER, J. ET AL. (2013): "GWAS of 126,559 Individuals Identifies Genetic Variants Associated with Educational Attainment," *Science*, 340, 1467–1471.

TERSKAYA, A. (2023): "Parental Human Capital Investment Responses to Children's Disability," *Journal of Human Capital*, 17.

TRAMPUSH, J. W., YANG, M. L. Z., YU, J. ET AL. (2017): "GWAS meta-analysis reveals novel loci and genetic correlates for general cognitive function: a report from the COGENT consortium," *Molecular Psychiatry*, 22, 336–345.

TREJO, S. AND DOMINGUE, B. W. (2018): "Genetic nature or genetic nurture? Introducing social genetic parameters to quantify bias in polygenic score analyses," *Biodemography and Social Biology*, 64, 187–215.

VILHJÁLMSSON, B. J., YANG, J., FINUCANE, H. K. ET AL. (2015): "Modeling linkage disequilibrium increases accuracy of polygenic risk scores," *The American Journal of Human Genetics*, 97, 576–592.

WRAY, N. R., YANG, J., HAYES, B. J. ET AL. (2013): "Pitfalls of predicting complex traits from SNPs," *Nature Reviews Genetics*, 14, 507–515.

YANG, J., LEE, S. H., GODDARD, M. E. ET AL. (2011): "GCTA: a tool for genome-wide complex trait analysis," *The American Journal of Human Genetics*, 88, 76–82.

Yengo, L., Sidari, M., Verweij, K. J. et al. (2020): "No evidence for social genetic effects or genetic similarity among friends beyond that due to population stratification: a reappraisal of Domingue et al (2018)," *Behavior Genetics*, 50, 67–71.

Young, A. I., Nehzati, S. M., Lee, C. et al. (2020): "Mendelian imputation of parental genotypes for genome-wide estimation of direct and indirect genetic effects," *BioRxiv*, 2020–07.

Zhu, Z., Bakshi, A., Vinkhuyzen, A. A. et al. (2015): "Dominance genetic variation contributes little to the missing heritability for human complex traits," *The American Journal of Human Genetics*, 96, 377–385.