

# POLYGENIC SCORE ANALYSIS OF EDUCATIONAL ACHIEVEMENT AND INTERGENERATIONAL MOBILITY

ALDO RUSTICHINI, WILLIAM G. IACONO, JAMES LEE, AND MATT MCGUE

**ABSTRACT.** A Genome-wide association study (*GWAS*) estimates size and significance of the effect of common genetic variants on a phenotype of interest. A Polygenic Score (*PGS*) is a score, computed for each individual, summarizing the expected value of a phenotype on the basis of the individual's genotype. The *PGS* is computed as a weighted sum of the values of the individual's genetic variants, using as weights the *GWAS* estimated coefficients from a training sample. Thus, *PGS* carries information on the genotype, and only on the genotype, of an individual. In our case phenotypes of interest are measures of educational achievement, such as having a college degree, or the education years, in a sample of approximately 2700 adult twins and their parents.

We set up the analysis in a standard model of optimal parental investment and intergenerational mobility, extended to include a fully specified genetic analysis of skill transmission, and show that the model's predictions on mobility differ substantially from those of the standard model. For instance, the coefficient of intergenerational income elasticity may be larger, and may differ across countries because the distribution of the genotype is different, completely independently of any difference in institution, technology or preferences.

We then study how much of the educational achievement is explained by the *PGS* for education, thus estimating how much of the variance of education can be explained by genetic factors alone. We find a substantial effect of *PGS* on performance in school, years of education and college.

Finally we study the channels between *PGS* and the educational achievement, distinguishing how much is due to cognitive skills and to personality traits. We show that the effect of *PGS* is substantially stronger on Intelligence than on other traits, like Constraint, which seem natural explanatory factors of educational success. For educational achievement, both cognitive and non cognitive skills are important, although the larger fraction of success is channeled by Intelligence.

---

*Date:* September 17, 2018.

We thank Aysu Okbay for generously running the meta-analysis (three times!) on the data, Peter Visscher for a clarification on Robinson et al. (2017), Philippe Köllinger for the help in the process, Joel Waldfogel, Tom Holmes, Giulio Zanella for very useful observations, criticisms and suggestions, and audiences in many seminars for very lively, illuminating discussions. Supported in part by grants from the National Science Foundation to AR (*SES1728056*), the National Institute on Alcohol Abuse and Alcoholism (*AA09367*) and the National Institute of Drug Abuse (*DA05147*).

## CONTENTS

1. Introduction	3
2. Parental investment and genetic transmission	5
2.1. Parental Investment	5
2.2. Skill Transmission	6
2.3. Matching Processes	8
2.4. Preferences and Stable Matchings	9
2.5. Complete Model	10
2.6. Intergenerational mobility in standard and genetic model	12
2.7. Correlation among Twins	15
2.8. Gene-Environment Correlation	17
2.9. Estimation Strategy	17
3. Methods	18
3.1. Computation of $PGS$	18
3.2. Measures of Educational achievement	18
3.3. Explanatory variables	19
4. $PGS$ and educational achievement	21
4.1. GPA score	21
4.2. College degree	22
4.3. Education Years	24
5. Identifying the path from $PGS$ to education	24
5.1. From Intelligence and Personality to education	25
5.2. From $PGS$ to personality	26
5.3. Mediation Analysis	28
6. Passive Gene Environment Correlation	31
7. Fixed Effects Analysis in $DZ$ twins	33
8. Conclusions	40
Appendix A. Appendix (not meant for publication)	43
A.1. Distribution of $PGS$	43
A.2. College achievement and $PGS$	44
A.3. Evidence of $rGE$	46
A.4. Regression Analysis	48
A.5. Assortative mating in Education	51
A.6. Evidence of Genetic Assortative Mating	52
A.7. Mediation Analysis	53
References	54

## 1. INTRODUCTION

In recent research of heritability of phenotypes based on genome-wide association studies (*GWAS*) a number of markers have been identified. A *GWAS* is a study of common genetic variants spanning the entire genome (typically one million Single Nucleotide Polymorphisms (*SNP*'s) or more) in a typically large set of individuals to determine if and how much any variant is associated with a trait. The markers that achieve significance at the conventional *GWAS* threshold<sup>1</sup> are still limited in number, and together explain a limited fraction of the variability of the phenotype. In spite of this, a considerable fraction of phenotypic variation can be explained by a larger set of genetic markers that includes variants not significantly associated with the phenotype.

A way to take into account the information available in markers, including perhaps those with significance lower than the *GWAS* threshold, is to compute a Polygenic Score (*PGS*). A *PGS* is an individual specific score, obtained as sum of the value of the markers in a selected set, each value weighted by a coefficient that has been estimated separately on an independent training sample (Dudbridge (2013)). Our analysis here is based on the large *GWAS* of educational attainment reported by Lee et al. (2018) (see also Rietveld et al. (2013), Okbay et al. (2016)). An illuminating discussion of the analysis of educational attainment in the modern *GWAS* era is in Cesarini and Visscher (2017).

We set up the investigation in a fully specified model of parental investment in education of children. The classical papers are Becker and Tomes (1979), Loury (1981), Becker and Tomes (1986). Important developments are, among many, in Solon (1992), Mulligan (1997), Mulligan (1999), Solon (2004), Black and Devereux (2011)). Our model differs from most existing ones in this field in two respects, both made necessary by the need to take into account the information on genotype and its transmission. First, we introduce explicitly the fact that children are the outcome of a joint process involving a father and a mother; so we need to include a theory of mating in the the model (similarly to Aiyagari et al. (2000), Greenwood et al. (2003)). The importance of assortative mating has been well documented in the past. For instance Greenwood et al. (2016) document that assortative mating along educational characteristics has increased in the USA. We build here on research like Fernandez and Rogerson (2001), Fernandez et al. (2005) which studies models where assortative mating directly affects intergenerational mobility. Second, we model the process of skill formation consistently with the transmission of genotype from parents to children, along well known lines in genetics (see for example Nagylaki (1992)).

Within this theoretical framework, we address two basic sets of questions. First, how much of the variance in educational achievement is explained by

---

<sup>1</sup>The threshold is  $5 \times 10^{-8}$ ; the factor  $10^{-8}$  corrects (Bonferroni) for multiple comparisons.

the *PGS*? Recalling that the score contains only genetic information, this estimate would give us a lower bound on how much of the variance of success in education can be attributed to the individual’s genotype. How is this effect mediated by assortative mating among parents, and the correlation among their genotypes? And finally, how is the effect of genes mediated by the direct effect on the genotype of the children, and how much mediated by the indirect effect on the environment provided to them, as well as parental investment?

Second, what are the channels through which the effect of genotype, as summarized by the *PGS* operates? Recall that the score is built on a statistical association between genotype and the phenotype of interest, in our case success in education. A natural first channel to consider is Intelligence: the score likely summarizes a set of highly polygenic effects on intelligence, and in turn intelligence improves the chances of success in education. But Intelligence is not the only plausible channel; personality traits are an important additional way. We use the term *personality* to indicate a set of individual characteristics possessed by a person that together determine a consistent pattern of cognition, emotions, motivations, and behaviors in various situations. A substantial fraction of success in education might be traced back to motivation, self-control, ambition; in general, personality traits distinct from pure cognitive skills. A gene affecting these traits would also appear as contribution to the *PGS* score, even if unrelated to intelligence. These are all natural channels. The effect of genes on education could operate, however, along completely different pathways, involving individual characteristics that have no bearing on the technology of educational attainment, for example discrimination. Clearly, understanding which of these pathways operates, and in what measure, is essential, particularly for policy guidance.

The paper is organized as follows. In section 2 we present the model, discuss its predictions, and how they differ from the standard model, particularly regarding inter-generational mobility. Data and methods used are reported in section 3. Section 4 argues that *PGS* is a good predictor of a substantial fraction of the variance in several measures of educational success at different age. Estimates of the model’s parameters predicting the effect from *PGS* to educational success are presented in section 5; different methods are used and compared. The effect of parental genotype on children’s success operating through the environment, in addition to the direct effect on their genotype (passive gene-environment correlation) is estimated in section 6. Fixed effects analysis of dizygotic (*DZ*) twins, in section 7, allows us to separate the role of environment (which is common for *DZ* twins) and genes (which are different in a fraction that we can estimate). Conclusions are presented in section 8.

## 2. PARENTAL INVESTMENT AND GENETIC TRANSMISSION

We begin by providing the conceptual and theoretical structure for our analysis below. To do so we must and do provide a model and an equilibrium concept. We build the model in section 2.1 to 2.4; the complete model to be tested is presented in section 2.5. Our aim is to show how the standard analysis of parental investment in education and inter-generational mobility (as pioneered in Becker and Tomes (1979), where the skill transmission follows a simple  $AR(1)$  process) should be modified to take into account a fully specified genetic mechanism of skill transmission. A core component of the model is the adaptation of the theory of marriage <sup>2</sup> (Becker (1973)) to predict mating and a model of genetic transmission. A comparison of the prediction of the two models is provided in section 2.6; we show that they differ substantially on key predictions, for instance on intergenerational mobility.

**2.1. Parental Investment.** A household maximizes a utility function of own consumption and future income of two children, which in turn is affected by the parental investment in education, genetic endowment and environment. The restriction to two children is consistent with the assumption that population size is constant. In our data, the two children also happen to be twins: this detail is irrelevant when we study parental investment, and only becomes important when we study the correlation of skill and income across siblings. We denote  $y$  the natural log of income,  $E$  consumption expenditure,  $I$  parental investment in education of children and  $h$  human capital measured by the education level.  $\epsilon^e$  and  $\epsilon^y$  denote the random shock to education and income: each one is *i.i.d.* across periods and the two are independent within periods.  $\alpha$ 's denote productivity parameters of the sub-scripted variable; so  $\alpha_I, \alpha_h$  denote positive real numbers.  $\delta \in (0, 1)$  is the discount factor. A vector of real numbers  $\theta = (\theta^1, \dots, \theta^{n_1}, \theta^{n_1+1}, \dots, \theta^n)$  describes the  $n$  skills, where index from 1 to  $n_1$  refers to hard or cognitive skills, and those from  $n_1 + 1$  to  $n$  to soft or non-cognitive skills (Heckman and Kautz (2012), Heckman and Kautz (2012)). Skills enter linearly into the production of the education level through an  $n$ -dimensional vector of coefficients  $\alpha_\theta$ . The superscript  $i$  refers to the family, the subscript  $j = 1, 2$  to the siblings; so a sibling is uniquely identified by the pair  $ij$ . Household log-income  $y^i$  is some combination of the log-income of father  $y_f^i$  and mother,  $y_m^i$  to be specified later.

---

<sup>2</sup>In this paper, two terms, matching and mating are used interchangeably, as synonymous for marriage. The reason for the multiple terms is that the term matching is used more frequently in the economics literature, and mating behavioral genetics. We use every time the term most appropriate in the context.

The  $i^{th}$  household solves:

$$(1) \quad \max_{(E^i, I_1^i, I_2^i)} E_{(\theta_1^i, \theta_2^i)} \left( (1 - \delta) \ln E^i + \delta \sum_{j=1,2} y_j^i \right),$$

subject to:

$$(2) \quad E^i + \sum_{j=1,2} I_j^i = Y^i$$

$$(3) \quad h_j^i = \alpha_I \ln I_j^i + \alpha_\theta \theta_j^i + \epsilon_j^{h,i}, j = 1, 2$$

$$(4) \quad y_j^i = \alpha_h h_j^i + \epsilon_j^{y,i}, j = 1, 2$$

We assume:

$$(5) \quad \forall i, j (\forall k \in \{h, y\} E \epsilon_j^{k,i} = 0); E \epsilon_j^{h,i} \epsilon_j^{y,i} = 0;$$

The choice on consumption and educational investment is taken with the knowledge of the skills  $(\theta_1^i, \theta_2^i)$  of the children, hence the sub-script in the expectation of equation (1), which refers to the random shocks  $\epsilon^h$  and  $\epsilon^y$ .

At the optimal solution of the problem in equations (1-5) optimal parental investment is equal for the two siblings ( $\hat{I}_1^i = \hat{I}_2^i \equiv \hat{I}^i$ ), and is a constant fraction of household income:

$$(6) \quad \hat{I}^i = \frac{\delta \alpha_{Ih}}{1 - \delta + 2\delta \alpha_{Ih}} \exp(y^i) \equiv \psi \exp(y^i).$$

where  $\alpha_{Ih} = \alpha_I \alpha_h$ . Equal investment in education for the two children is if course a very special feature due to the preferences we have adopted.

**2.2. Skill Transmission.** We replace the standard  $AR(1)$  mechanism of skill transmission (discussed in section 2.6 below) with a detailed model where the skill vector  $\theta$  results from genetic factors, parental investment in education, family environment common to all children, and idiosyncratic random events for each individual. We examine these components separately, beginning with the genetic component.

If  $K$  is the number of loci, a genotype is a  $g \in G^K \equiv \{0, 1, 2\}^K$ , where 0, 1, 2 refers to the count of one of the alleles in a bi-allelic system (*GWAS*'s overwhelmingly deal with variants, *SNP*'s, that are bi-allelic in the analysis). The joint distribution of genotypes of the children, given the genotype of the two parents, depends on the twin type, that may be monozygotic, *MZ* or dizygotic, *DZ*. To define it we start with the function from parents' genotype to the probability over genotypes of an individual offspring, described by a function  $H$  from  $G^K \times G^K$  to  $\Delta(G^K)$ :

$$(7) \quad H : (g_m, g_f) \mapsto H(g_m, g_f) \in \Delta(G^K).$$

$H$  follows well known rules of Mendelian inheritance; for instance if  $K = 1$  so  $G^K = \{0, 1, 2\}$ , then  $H(1, 1)$  is (0.25, 0.5, 0.25), and  $H(0, 2)$  is (0, 1, 0).

Children in our sample are all twins; in this case the genetic transmission functions depend on the twin type  $T \in \{DZ, MZ\}$ , and are defined as:

$$(8) \quad H_{DZ}(g_m, g_f)(g^1, g^2) = H(g_m, g_f)(g^1)H(g_m, g_f)(g^2)$$

for the genotype pair  $(g^1, g^2)$  of the  $DZ$  twins and

$$(9) \quad \begin{aligned} H_{MZ}(g_m, g_f)(g^1, g^2) &= H(g_m, g_f)(g^1) \text{ if } g^1 = g^2 \\ &= 0 \text{ otherwise} \end{aligned}$$

for  $MZ$  twins.

Let  $w$  denote the  $n$ -valued function determining skills by the genotype  $g$ . The use of polygenic scores relies on the two assumptions that  $w$  is additive across loci—that is,

$$(10) \quad w(g) = \sum_{k=1}^K w_k(g_k)$$

and within each locus—that is,

$$(11) \quad \forall k, w_k(g_k) = \beta_k g_k \text{ for some real number } \beta_k$$

$X$  a vector of observable variables (which include for instance the parents' education, the family income and social status, and so on),  $\Pi$  a matrix with  $n$  rows,  $F$  a family specific  $n$ -dimensional shock (common to both twins, either  $MZ$  or  $DZ$ ), and  $\epsilon^\theta$  an individual specific  $n$ -dimensional environmental zero-mean shock on the skill. The skill of twin  $ij$  is given by:

$$(12) \quad \theta_j^i = w(g_j^i) + \Pi X_j^i + F^i + \epsilon_j^{\theta, i}.$$

In the analysis below we pay special attention to the simplified model where the only observable variable in the vector  $X$  in equation (12) is parental income, modeling the effect of income on skill, distinct from the effect of parental investment on human capital in equation (3). Equation (12) becomes:

$$(13) \quad \theta_j^i = w(g_j^i) + \Pi y^i + F^i + \epsilon_j^{\theta, i}$$

We assume:

$$(14) \quad \forall i, j (\forall k \in \{h, y\} E \epsilon_j^{k, i} \epsilon_j^{\theta, i} = 0), EF^i (\epsilon_j^{\theta, i})^T = 0.$$

In the analysis below we also use the more general model to control for education of parents, college degree of parents, work status of the father. Substituting the optimal investment determined by equation (6) into (3) and substituting the result into equation (4) we get the reduced equation for income:

$$(15) \quad y_j^i = a + \alpha_{Ih} y^i + \alpha_{\theta h} \theta_j^i + \alpha_h \epsilon_j^{h, i} + \epsilon_j^{y, i}$$

where  $a = \alpha_{Ih} \ln \psi$ , and  $\alpha_{\theta h} = \alpha_\theta \alpha_h$ .

The complete model of the process on genotype, income, education and skill for twins of type  $T$  is given by equations (3) for education, (8) and (9)

for the genotype transmission, (12) for skill, and the above reduced equation (15) for income. In the simple model (13) we assume

$$(16) \quad \alpha_{Ih} + \alpha_{\theta h} \pi < 1$$

These equations completely determine a non-linear (because of the function  $H$  in equation (7)) transition on measures on the space of genotypes and income,  $\Delta(G^K \times \Theta \times Y)$  when we specify how the pairs of parents are selected. To this we turn now.

**2.3. Matching Processes.** To complete the system (8, 9, 12, 15) we need to specify the matching process for parents. We assume that this process depends on the individual characteristics that we have described so far, skill and income, which are relevant for economic outcomes, but also on characteristics in a set  $C$  that are important for matching but not for economic activity (such as, to a first approximation, physical appearance); we let  $Z \equiv G^K \times \Theta \times Y \times C$ , and the observable characteristics  $Z_O \equiv \Theta \times Y \times C$  with generic element  $z_O$ ; for convenience we indicate with a subscript whether the element in  $\Delta(Z)$  refers to the mother (as in  $\Delta_m(Z)$ ) or the father.

A *matching* associates to a pair of distributions  $(\mu_m, \mu_f) \in \Delta_m(Z) \times \Delta_f(Z)$  an element  $M(\mu_m, \mu_f) \equiv \nu \in \Delta(Z \times Z)$ , describing the distribution of pairs of genotypes, skills, income and characteristics of the two parents. The matching process is required to be:

- (1) **Feasible:** the marginal over the parents distribution is the same as the original one:

$$M(\mu_m, \mu_f)_{\Delta_i(Z)} = \mu_i, i \in \{m, f\}$$

- (2) **Independent of genotype:** the matching only depends on the observable characteristics  $z_O \in \Theta \times Y \times C$ .

The independence assumption is reasonable, at least as long as genetic testing has not become widespread: it assures that for every  $z_{mO}$ , and  $z_{fO}, z'_{fO}$ , for every pairs of genotypes  $(g_m, g_f)$  and  $(g'_m, g'_f)$  the derivative below only depends on  $z_{mO}, z_{fO}, z'_{fO}$

$$(17) \quad \frac{d\nu(g_m, z_{mO}, g_f, z_{fO})}{d\nu(g'_m, z_{mO}, g'_f, z'_{fO})} = R(z_{mO}, z_{fO}, z'_{fO})$$

Random Matching is an example of matching. With random matching, a mother of type  $z_{mO}$  is selected, and independently a  $z_{fO}$  for the father, according to  $\mu_m$  and  $\mu_f$  respectively. This model is convenient for its simplicity, but it is not supported by the data, which show instead substantial positive correlation between several characteristics of the parents. Thus a model induced by preferences over matchings is desirable, and a better approximation.



**2.4. Preferences and Stable Matchings.** We assume a preference order over matchings; consistently with our assumption on matchings, the order is defined on the observable vector  $z_O$  of each of the two mates. It is also monotonic in the  $\Theta \times Y$  component, and is homophilic in the  $C$  component. More precisely, recall that  $\Theta \equiv \times_{l=1}^n \Theta_l$ , each component has a natural order (taller, more intelligent, lower Neuroticism score), and  $Y$  has a natural order, so  $\Theta$  and  $\Theta \times Y$  have the natural induced partial order. An *individual* in the marriage market is a type  $z_O \in \Theta \times Y \times C$ . Preferences over mates of the individual  $z_O$  of sex  $s \in \{m, f\}$  (recall  $m$  is mother, assumed to be female) are represented by a weak order  $\succeq_{z_{sO}}$  that is monotonic:

$$(18) \quad \forall z''_M, z'_M : z''_M \geq z'_M \text{ implies } \forall c \in C, (z''_M, c) \succeq_{z_{sO}} (z'_M, c)$$

and homophilic:

$$(19) \quad \forall z_M, c, e, f : d(f, c) \leq d(e, c) \text{ implies } \forall z'_M (z'_M, f) \succeq_{(z_M, c)_s} (z'_M, e).$$

The household maximization problem described in equation (1)-(5), which only depends on the  $\Theta \times Y$  components, defines a preference over matches. In the household maximization problem an individual  $(\theta_m, y_m)$  evaluates the utility  $U(\theta_m, y_m, \theta_f, y_f)$  from a match with an individual  $(\theta_f, y_f)$  anticipating the household income and the skill of the two children; so her preferences (if the preferences are completely described by the household maximization problem) are represented by  $U(\theta_m, y_m, \cdot)$ . The same holds for the  $f$  potential spouse. We assume that household log income  $y^h$  is linear combination of the income of the two spouses with weights  $w^{y_i}$  adding to 1, and that the expected (by the parents) skill of each child  $\theta^c$  is linear combination of the skills of the parents with weight  $w_i^\theta$ ,  $i \in \{m, f\}$  also adding to 1. In summary we assume:

$$(20) \quad y^h = w_m^y y_m + w_f^y y_f;$$

and

$$(21) \quad \theta^c = w_m^\theta \theta_m + w_f^\theta \theta_f;$$

Substituting the optimal investment (6) into the budget constraint (2)), the education (3) and income(4) equations we find that, up to a constant independent of  $\theta$  and  $y$ , the *worth* in the marriage market of a type  $(\theta, y)$  of sex  $i \in \{m, f\}$  is:

$$(22) \quad W_i(\theta, y) \equiv (1 - \delta + 2\delta\alpha_{hI})w_i^y y + 2\delta\alpha_{h\theta}w_i^\theta \theta$$

and the utility of a household is the sum of the worth of the spouses:

$$(23) \quad U(\theta_m, y_m, \theta_f, y_f) = W_m(\theta_m, y_m) + W_f(\theta_f, y_f)$$

so the household utility from the households maximization problem is linear and monotonically increasing in the parents' types and incomes, hence the overall utility is (if we assume that any additional components are monotonically increasing ) monotonically increasing.

A *stable matching* is defined as usual a matching that cannot be blocked by individuals or pairs of mates.<sup>3</sup> By the properties we have derived we conclude:

**Proposition 2.1.** *A stable matching exists, and the distribution over genes conditional on any pair of observable characteristics is uniform, that is (17) holds.*

Note that the stable matching does not in general imply perfect segregation according to characteristics  $C$ ; imperfect segregation may occur if the distribution of the  $(\theta, y)$  components is not the same conditional on the  $C$  characteristics.

**2.5. Complete Model.** The genotype and income process determine completely the equilibrium. Note first that:

**Proposition 2.2.** *If assumption (16) holds the system has an invariant distribution  $\mu \in \Delta(Z)$ .*

We can then subtract from the variables  $(y_j^i, \theta_j^i, h_j^i, w(g_j^i))$  their expected value with respect to the invariant distribution; so the constants are eliminated (for example the  $a$  term in the reduced equation for income is eliminated). Since no confusion is possible, we keep the same names for these variables which have now zero mean. We write the equations (8) and (9) in the compact form:

$$(24) \quad g_j^i \sim H_T(g_m^i, g_f^i), \quad T \in \{MZ, DZ\}.$$

If we substitute equation (13) into the reduced equation for income (15) we get the twin's income as function of genetic endowment, family income, family environment and idiosyncratic shocks:

$$(25) \quad y_j^i = \alpha_{\theta h} w(g_j^i) + (\alpha_{Ih} + \alpha_{\theta h} \pi) y^i + \alpha_{\theta h} F^i + \alpha_{\theta h} \epsilon_j^{\theta, i} + \alpha_h \epsilon_j^{h, i} + \epsilon_j^{y, i}.$$

Our *PGS* measure is a summary of the effect of  $w(g_j^i)$ . The analysis of the invariant distribution is simple if we set

$$(30) \quad w_m^\theta = w_f^\theta, w_m^y = w_f^y,^4$$

---

<sup>3</sup>More precisely: A matching  $\nu$  is stable, if and only if for all, except possibly a zero measure set (with respect to the product measure  $\nu \otimes \nu$ ), pairs  $(z_m, z_f, z'_m, z'_f)$ ,

$$z_f \succ_{z_m} z'_f \text{ or } z'_m \succ_{z'_f} z_m \text{ or } \left( z_f \succeq_{z_m} z'_f \text{ \& } z'_m \succeq_{z'_f} z_m \right).$$

<sup>4</sup>The system resulting if we assume equation (26) is:

$$(27) \quad y_j^i = \alpha_{\theta h} \theta_j^i + \alpha_{Ih} y_j^i + \epsilon_j^{y, i}$$

$$(28) \quad g_j^i \text{ distributed as } H(g_m^i, g_f^i)$$

$$\Pi = 0, F^i = 0, \epsilon^\theta = 0, \epsilon^h = 0, \epsilon^y \sim N(0, \sigma_{\epsilon^y}^2).$$

We call *worth class* the set of individuals with the same worth. In this model, in each generation children are born of spouses of same worth (not necessarily income: higher skill may compensate a lower income).

We call the *skill allele* at a locus the allele which yields a higher value of the skill (more precisely, it has a higher genic value).<sup>5</sup> At equilibrium, matching is random within each worth class, thus alleles are in Hardy-Weinberg equilibrium. But the frequency may differ across classes:

**Proposition 2.3.** *Assume that skill affects human capital formation ( $\alpha_{\theta h} \neq 0$ ), and that the worth of an individual depends on income and skill, as in equation (30). Then alleles at each locus are in Hardy-Weinberg equilibrium within each worth class. Also the frequency of the skill allele is increasing with the worth, thus society is stratified according to income, with higher frequency of skill alleles associated with higher income.*

Figure 2 below illustrates the second part of the proposition.

**2.5.1. Numerical computation.** In this simple case the main properties of the process and equilibrium distribution of the model can be illustrated with a numerical computation of the equilibrium.<sup>6</sup> We study the distribution in  $\Delta(Y \times G^K)$  in successive generations of a constant size population where each household has two children, and a single skill ( $n = 1$ ). The sex of each child is determined independently (from each other and from the other variables) with probability 1/2 on each sex.

*Speed of Convergence.* Convergence to the invariant distribution is fast, and approximately achieved with five generations. Figure 1 reports the value, for each  $k$  generation, of the ratio of the norm of the difference between current and past  $\mu$ , and the norm of the current  $\mu$ . The ratio is within ten per cent after five generations.

*Stratification.* The skill allele has at equilibrium a frequency that is increasing with worth, but also with income. As we mentioned in proposition 2.3, society is stratified. The effect is strong, and is stronger the higher the genic value of the allele. Both facts are illustrated in Figure 2.

---

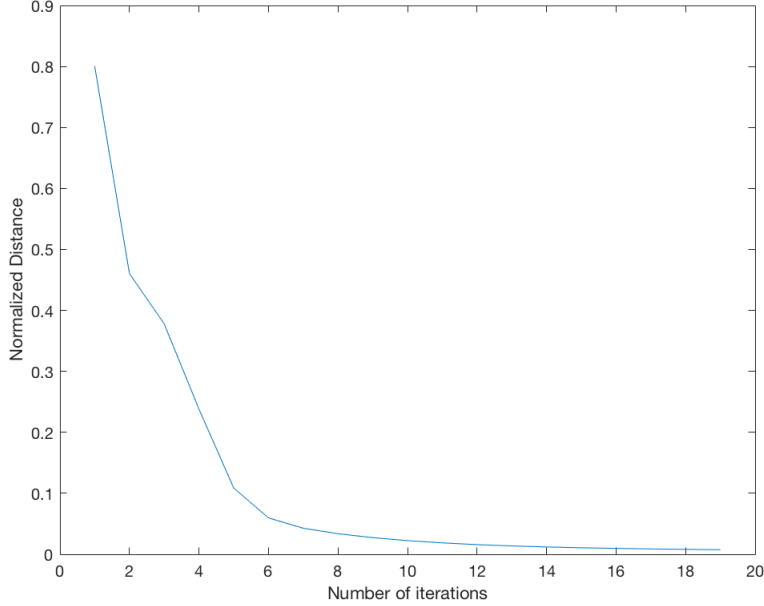

$$(29) \quad \theta_j^i = w(g_j^i)$$

$$(30) \quad W(\theta, y) = (1 - \delta + 2\delta\alpha_{Ih})y + 2\delta\alpha_{\theta h}$$

<sup>5</sup>The genic value is a measure of the contribution of the allele to the phenotype of interest, the skill in our case (see for example page 117 of Crow and Kimura (1970)).

<sup>6</sup>Coding in Matlab (R2017a). The Matlab code is available upon request.

FIGURE 1. **Speed of convergence to invariant distribution.** *Normalized distance* on the vertical axis is the ratio of the norm of the difference between the current and previous distribution, and the norm of the current distribution.



**2.6. Intergenerational mobility in standard and genetic model.** The standard model with autoregressive transmission of skill (Becker and Tomes (1979)) has (in our notation) the following equations for income in generation  $t$ :

$$(31) \quad y_{t+1} = \alpha_{Ih}y_t + \alpha_{\theta h}\theta_{t+1} + \epsilon_{t+1}^y$$

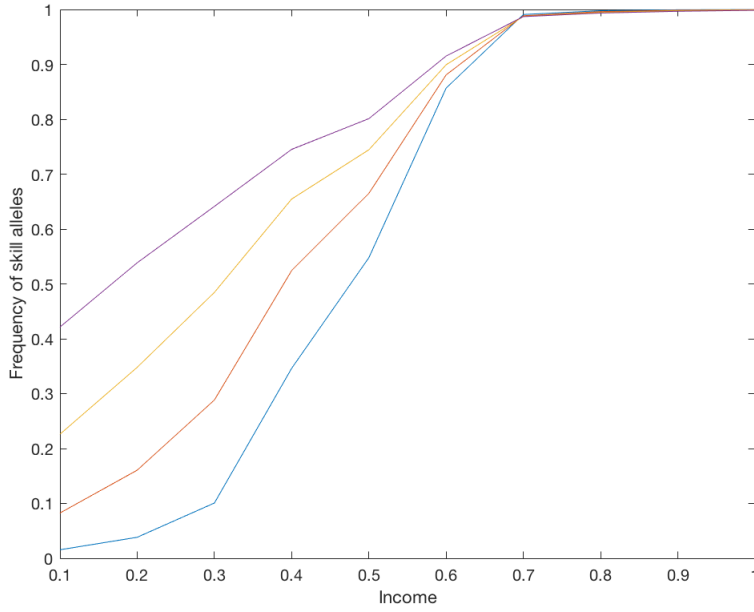
and for skill:

$$(32) \quad \theta_{t+1} = \eta\theta_t + \epsilon_{t+1}^\theta$$

where  $\eta \in (0, 1)$  is a fixed “heritability” parameter. Note that there is only one type of skill. At the stationary distribution, we can compute, using the Yule-Walker equations, the intergenerational income elasticity  $\rho_{PM}$  (the reason for the subscript  $PM$  will be soon clear) to be:

$$(33) \quad \rho_{PM} = \alpha_{Ih} + \alpha_{\theta h} \frac{\eta E(\theta, y)}{Var(y)}$$

FIGURE 2. **Frequency of the skill allele by income.** The skill allele is the one higher genic value (relatively larger effect on skill). The blue line, with largest differences across income, describes the frequency of the allele with highest genic value; the others are in decreasing order.



where  $E(\theta, y)$  and  $Var(y)$  have an explicit expression in terms of the primitive parameters.<sup>7</sup> When  $\sigma_{\epsilon y} = 0$ , the inter-generational persistence formula (33) becomes the well known formula (see e.g. Solon (2004)) where persistence is a simple weighted average of the income and skill transmission:

$$(37) \quad \rho_{PM} = \frac{\alpha_{Ih} + \eta}{1 + \alpha_{Ih}\eta}$$

---

<sup>7</sup>The explicit expressions are:

$$(34) \quad V(\theta) = \frac{\sigma_{\epsilon\theta}^2}{1 - \eta^2}$$

$$(35) \quad E(\theta, y) = \frac{\alpha_{\theta h} V(\theta)}{1 - \alpha_{Ih}\eta}$$

$$(36) \quad V(y) = \frac{1}{1 - \alpha_{Ih}^2} (\alpha_{\theta h}^2 V(\theta) + \sigma_{\epsilon y}^2 + 2\alpha_{Ih}\alpha_{\theta h}\eta E(\theta, y))$$

A direct comparison of the standard model (equations (31) and (32)) with a genetic model like (24) and (25), where sex is an essential component of reproduction, is meaningless, since, apart from the genes, there are not even two parents in the standard model. So we must first build a more general model which includes the standard one as a special case of the general class of models (with gametic reproduction, as is the case for human population) in sections 2.5 and 2.4. We assume income and skill to be the weighted average of the income and skill of the two parents, as in equations (20) and (21). Thus, the income of the child follows the equation:

$$(38) \quad y_{t+1} = \alpha_{Ih} \sum_{i=m,f} w_i^y y_{it} + \alpha_{\theta h} \theta_{t+1} + \epsilon_{t+1}^y$$

and the skill transmission follows:

$$(39) \quad \theta_{t+1} = \eta \sum_{i=m,f} w_i^\theta \theta_{it} + \epsilon_{t+1}^\theta$$

The matching between parents that decides the pairing of  $(\theta_{mt}, y_{mt})$  with  $(\theta_{ft}, y_{ft})$  is determined by preferences and stable matching as in section 2.4. The standard model (31 - 32) becomes a special case of (38 - 39) when we assume that preferences of mothers and fathers are lexicographic (with any order on  $\theta$  and  $y$ ) and  $\mu_m = \mu_f$ , so matching occurs only among identical types (Perfect Matching, hence the *PM* subscript).

We now show that the formulas for intergenerational income elasticity (33) or (37) of the standard model are an *upper bound* on the persistence within the class of models requiring equations (31), (38) and (39). The reason is that, as we have just seen, the standard model maximizes the similarity among parents, forcing their income and skill to be identical. For example consider the case where parents match only on income, but may differ in skill. This happens when preferences are linearly ordered by the income of the spouse. In this case, the corresponding intergenerational elasticity, call it  $\rho_{MY}$ , can be shown to satisfy: <sup>8</sup>

$$(41) \quad \rho_{MY} < \rho_{PM}$$

We can now discuss the relation between prediction of the standard and genetic model on the important issue of the size of intergenerational mobility. The standard model with autoregressive transmission of skill assumes a fixed  $\eta$  (in equation 32). Such a fixed parameter, however, does not exist: the genetic model shows that the persistence represented by that  $\eta$  is endogenous, and depends on the distribution of the genotype. Therefore

---

<sup>8</sup>The inequality follows because when parents match on income and only on income the system (34-36) is as follows. Equation 34 becomes:

$$(40) \quad V(\theta) = \frac{\sigma_{\epsilon^\theta}^2 + \frac{\eta^2}{2} E(\theta_m, \theta_f)}{1 - \frac{\eta^2}{2}}.$$

Equations (35) and (36) are unchanged. Rearranging one obtains the inequality 41.

the corresponding elasticity, call it  $\rho_G$ , also does depend on the distribution, which is different in different populations. So persistence may differ among populations independently of preferences, technology and institutions in the economy, but depending only on the distribution of the genotype in that population.

An important implication of the differences we have highlighted so far is that the persistence in a model with genetic transmission of skill can be *higher* than the one in the the standard model, even higher than the highest possible value in the class of standard models with sexual reproduction (equations 38 and 39). That is, it may be the case that  $\rho_G > \rho_{PM}$ . It follows in particular that the adoption of the amended model with  $AR(1)$  transmission and sexual reproduction (equations (38 –39)) might make predictions worse, by further underestimating the persistence.

We illustrate this possibility in a simple but clarifying example. Take  $K = 1$  (a single locus with alleles  $\{A, a\}$ ), with frequency  $p(A)$  of  $A$ , determining a one dimensional skill  $\theta \in \{\theta_0, \theta_1, \theta_2\}$ , ordered as the index. Preferences are determined by the household maximization problem, hence are described by (23); and to ease comparison with the simple form (37) we assume  $\sigma_{\epsilon^y} = 0, \Pi = 0, F = 0, \epsilon^\theta = 0$ .

This economy has a stationary distribution at two values:

$$(42) \quad (0, y_0, \theta_0) \text{ with prob } 1 - p(A), (2, y_2, \theta_2) \text{ with prob } p(A),$$

where

$$y_i = \frac{\alpha_{\theta h} \theta_i}{1 - \alpha_{Ih}}.$$

The persistence here is 1, and this can never occur in an autoregressive model with  $\eta < 1$ .

The example is obviously artificial in the assumption that a skill phenotype is determined by a single locus, whereas the skills of interest for economic applications are highly polygenic; the force highlighted by the example, however, is not at all artificial, and points to the effect that assortative mating has on pulling apart the distribution of the genotype into classes of homozygous individuals.<sup>9</sup> This effect is absent by assumption in the autoregressive model, even in the amended version given by equations (31), (38) and (39).

**2.7. Correlation among Twins.** In the fixed effects analysis below we rely on the fact that  $DZ$  twins share important environmental characteristics, but do *not* entirely share the genotype. The degree of the sharing depends on the nature and strength of the assortative matching between parents. Genetic correlation among parents may occur for two different

<sup>9</sup>This force is well recognized in population genetics: see chapter 4 of Crow and Kimura (1970), in particular sections 4.6 for our single locus example and 4.7. for a multifactorial. The analysis in population genetics is very different from the one we present here because the assortative mating in our model is endogenous and determined at equilibrium in the marriage market.

reasons. Correlation may exist because matching is directly on the relevant phenotype (for example, the correlation on genes affecting Intelligence among parents occurs because parents match according to Intelligence); or it may occur indirectly, when matching occurs along dimensions unrelated to the phenotype (for example, matching occurs along the characteristics in the set  $C$  of physical appearance), but due to population stratification a correlation between genes affecting variables in  $C$  and  $\Theta$  exist. We can illustrate this second possibility considering the extreme case in which there is no overlap between loci affecting the  $\theta$  skills and the characteristics in  $C$ , and matching along  $C$  characteristics is perfect. In this case the stationary distribution has segregated populations with different frequencies on the alleles determining  $\theta$ , thus different distributions on the  $\theta$  skills. This equilibrium is not robust, of course: with a small imperfection in the  $C$ -matching the frequency of the  $\theta$  alleles converges exponentially in the long run to a value independent of the  $C$  characteristics; however, the transition is slow when the imperfection is small and in the transition the correlation may be substantial.

Whatever the cause, the correlation for  $DZ$  twins is a simple function of the correlation between the  $PGS$  of the parents. We use subscripts 1, 2 and  $m$  and  $f$  to indicate that the variable refers to first and second sibling, mother and father respectively, and indicate with  $\mathcal{G}_i$  be the algebra of genotype of  $i$ , with  $i = m, f$ . Then:

**Lemma 2.4.** *The correlation between the  $PGS$  of non identical full siblings, hence in particular of  $DZ$  twins, is equal to  $\frac{1}{2}$  plus half of the correlation between  $PGS$  of the parents, that is:*

$$E(PGS^1 PGS^2) = \frac{1}{2} + \frac{1}{2}E(PGS_m PGS_f)$$

*Proof.* The proof follows well known lines (Nagylaki (1992)); they are perhaps less familiar to an audience of economists, so it is presented here.

$$\begin{aligned} & E(PGS_1 PGS_2) \\ &= E(E(PGS_1 PGS_2) | \mathcal{G}_m, \mathcal{G}_f) \\ &= E(E(PGS_1) | \mathcal{G}_m, \mathcal{G}_f) E(PGS_2) | \mathcal{G}_m, \mathcal{G}_f) \\ &= E(E(\frac{1}{2}(PGS_m + PGS_f)) | \mathcal{G}_m, \mathcal{G}_f) E(\frac{1}{2}(PGS_m + PGS_f)) | \mathcal{G}_m, \mathcal{G}_f) \\ &= \frac{1}{2} E(E((PGS_m)^2 + PGS_m PGS_f) | \mathcal{G}_m, \mathcal{G}_f) \\ &= \frac{1}{2} + \frac{1}{2} E(PGS_m PGS_f) \end{aligned}$$

where the first equality follows from elementary property of expectation, the second from the conditional independence of  $PGS$  with respect to parents' genotype, the third from additivity of  $PGS$  of each offspring (our definition of  $PGS$  ignores the dominance effects), the fourth from symmetry between  $PGS_m$  and  $PGS_f$ , and fifth follow from elementary properties of expectations.  $\square$

Lemma 2.4 gives the correlation among  $DZ$  twins as a function of the correlation among parents. In section 7 below we estimate the correlation



among parents' *PGS*, and find a perfect match with the prediction of the lemma.

**2.8. Gene-Environment Correlation.** Genes and environment operate together to determine personality and behavior of individuals. Importantly, choice of environment is included in behavior. Behavior genetic research in the last forty years has refined the conceptual tools to model the way in which this joint effects operate (Plomin et al. (1977), Scarr and McCartney (1983), Jaffee and Price (2007)). *Gene-environment interaction* (usually denoted  $G \times E$ ) describes the idea that even if genes and environment are independent, the way in which each of the two operates on personality and behavior may depend on the value of the other; that is, genes and environment do not operate additively. For example, genes may determine the motivation of an individual (as a personality trait, measured for example by tasks or survey questions) and environment may offer opportunities (measured for instance by schooling available in the place of origin); but the resulting success of the individual (measured by education or income) may be different from a linear combination of the two. In an extreme example, in poor environments where opportunities are severely constrained a person with high motivation and intelligence may fail just as one with low values, and the difference may emerge only when adequate opportunities are offered.

The idea of *Gene-Environment correlation* (usually denoted as  $rGE$ , although sometimes (Scarr and McCartney (1983)) the more explicit and restrictive notation  $\text{Gene} \rightarrow \text{Environment}$  is used) rejects the assumption that environment and genes are uncorrelated. There are three main ways in which the correlation may arise. The most important for our purposes is the *passive rGE* effect.<sup>10</sup> Genes of the parents affect directly the genes of the children; but they also affect the environment in which the child grows, hence the potential for correlation between the two. Of course, the action of the genes of the parents on the environment does not arise directly as an effect of genes: Higher Intelligence of parents, due in part to the genes of the parents, affects directly the genes of the children, and together with shared environment, parental investment on the parents, affect the environment of the children.

**2.9. Estimation Strategy.** The next sections test and estimate the parameters of the model given by equations (13), (24) and (25). The data we have available for income of twins are at the moment limited to income at age 29 (see section 3.3 for details), hence they reflect only approximately the full earnings potential, and are therefore not yet suitable for a test. We

---

<sup>10</sup>The other two effects are *evocative* and *active*. The evocative effect refers to the difference in response that different genotypes induce in the environment; for instance, more active children are more likely to induce stronger social stimulation from the environment, and hence richer learning. The active effect is produced by the selection, perhaps purposeful, of different environment by different genetic types. These two effects are harder to measure in our data than the effects of passive  $GE$  correlation.

have instead reliable data on educational achievement, and in our model income is a linear function of human capital (see equation (4)), so we focus in particular on the model (13), (24) and (43). The latter equation describes human capital accumulation:

$$(43) \quad h_j^i = \alpha_I y^i + \alpha_\theta \theta_j^i + \epsilon_j^{h,i}$$

and is obtained substituting (6) into (3) and subtracting the constant term. We are in particular interested in an estimate of  $w$ , the function mapping from genotype (measured by the *PGS*) and skills, and the parameters  $\alpha_\theta$  affecting the different components of personality. This analysis is presented in section 4; the interpretation of the estimated coefficients and the possible pathways is presented in section 5.

In our model, the passive correlation is modeled as the effect of parents' genotype on the characteristics affecting the environment of the children. As clear from the model (using equation (25) for both parents) the correlation is mediated by the environmental effect (in particular the income of the household of the *parents*, not just the parents' genes). Results of the analysis of the passive *rGE* are presented in section 6.

As we mentioned in the remarks after lemma 2.4 the bound on the correlation among *DZ* twins allows us to use fixed effects analysis of the role of *PGS*. In section 7 below we estimate the correlation among parents' *PGS*, and the results of the fixed effects analysis. We can now proceed with the description of the data.

### 3. METHODS

**3.1. Computation of *PGS*.** The *PGS*'s of subjects were computed using *LDpred*, (Vilhjlmsson et al. (2015)), with prior probability of 0.3, using weights for each single nucleotide polymorphism (*SNP*) estimated from the meta-analysis in Lee et al. (2018), excluding our (*MTFS*, described below) sample. The analysis is restricted to individuals of European descent. We control for Principal components (the potential problems associated with population stratification have been recently highlighted in Sohail et al. (2018) and Berg et al. (2018)).

**3.2. Measures of Educational achievement.** Individuals in the sample we use here are twin participants in the Minnesota Twin Family Study (MTFS) (Iacono et al. (1999), Disney et al. (1999)), which includes two cohorts of twins, one assessed initially at a target age of 11 (N=1512) and a second assessed initially at a target age of 17 (N=1252), and subsequent follow-up assessments undertaken at target ages of 20, 24 and 29 for the older cohort and 14, 17, 20, 24 and 29 for the younger cohort. The participation rates in the follow-ups of MTFS have generally been above 90 % McGue et al. (2014). Information on educational achievement in the sample is provided by a classification of the individual in seven classes, described in Table 1.

Data on academic performance of the twins in school were collected in a dedicated academic history interview, given to both mother and child. Four scores were calculated: GPA, Behavior Problems, Academic Problems and Academic Motivation.

The *GPA score* used here is a GPA-like index, not the actual *GPA*. Five questions in the Academic History survey asked separately both the mother and the child about grades the child was getting in school. The questions provided a 5-point letters scale, from *A* to *F* for the answer. The questions asked about grades in (a) Reading/English, (b) Arithmetic/Math, (c) Science, (d) Social Studies/History, and (e) Overall. The *GPA* score was then calculated to represent an average of items *a* – *d* transformed to a four-point scale. In a validation sample (Johnson et al. (2004)), the correlation between reported grades and actual GPA from school transcripts exceeded 0.80.

TABLE 1. Education years variable. The variable Class is a coarser classification used in the analysis.

Education level	Class	Years
less than HS	1	10
GED	1	11
HS	2	13
HS + Vocation	3	14
Community college	3	15
College	4	19
Professional degree	5	22

**3.3. Explanatory variables.** A specific strength of our data is the availability of information on variables that are natural candidates to provide an explanation of the way in which the genetic profile of individuals, summarized by the PGS, can affect educational achievement. We describe these data here.

*Cognitive ability.* Cognitive ability was assessed at intake for both MTFs cohorts using four subtests from the age-appropriate Wechsler Intelligence Scale. Twins in the younger cohort were assessed with the Wechsler Intelligence Scale for Children-Revised (WISC-R) and twins in the older cohort were assessed with the Wechsler Adult Intelligence Scale-Revised (WAIS-R). The short forms consisted of two Performance subtests (Block Design and Picture Arrangement) and two Verbal subtests (Information and Vocabulary), and the scaled scores from these subtests were prorated to determine overall IQ. IQ from this short form has been shown to correlate ( $r = 0.94$ ) with IQ from the complete test (Sattler (1974)).

*Non-cognitive Skills: Personality measures.* Six measures of non-cognitive skills derived from the age-17 assessment of both cohorts were used. First, we used three higher-order scales from the Multidimensional Personality Questionnaire (MPQ, Tellegen and Waller (2008)). The *MPQ* has eleven primary trait scales (Absorption, Well-Being, Social Potency, Achievement, Social Closeness, Stress reaction, Aggression, Alienation, Control, Harm Avoidance, Traditionalism). Each is assessed with 18 self-report items. The three higher order *MPQ* scales (Positive Emotionality (here PE, associated with Wellbeing, Social Potency, Achievement, and Social Closeness), Negative Emotionality (NE, associated with Stress Reaction, Alienation, and Aggression) and Constraint (CN, associated with Control, Harm Avoidance, and Traditionalism.)) are computed as linear functions of the 11 primary scales.<sup>11</sup>

High Constraint is associated with tendencies to inhibit and constrain impulsive as well as risk-taking behavior. Individuals with higher Negative Emotionality scores are more prone to experience anxiety anger, and in general negative engagement. Positive Emotionality is associated with search for rewarding behavior and experience, while low PE may be associated with loss of interest, depressive engagement and fatigue. In our sample the three higher order dimensions, as well as IQ, are approximately normally distributed.

*Additional Non-cognitive Skills.* Three additional measures of soft skills were derived from answers to questionnaires.

*Externalizing* was the total number of DSM-IV symptoms of oppositional defiant disorder, conduct disorder, and adult antisocial behavior (i.e., the adult symptoms used in diagnosing antisocial personality disorder) obtained by interviewing the twin using with the Diagnostic Interview for Children and Adolescents (DICA-R) (Reich (2000), Welner et al. (1987)) and Structured Clinical Interview for DSM-III-R (SCID) Spitzer et al. (1992)). The interviews were modified to ensure complete coverage DSM-IV and symptoms were reported over the lifetime of the adolescent. In the analysis reported here, the Externalizing scale was log-transformed (after adding 1) to minimize positive skew.

The *Academic Effort* scale consisted of eight items answered by the twins' mother on a 4-point scale (Definitely False, Probably False, Probably True, Definitely True). Items on this scale ( $\alpha = 0.91$ , (Cronbach (1951))) cover academic effort (e.g., "Turns in homework on time") and motivation ("Wants to earn good grades").

Finally, the *Academic Problems* scale consisted of three items ( $\alpha = .77$ ) answered on the same 4-point format by the mother and covering behavioral problems in a school setting (e.g., "Easily distracted in class").

---

<sup>11</sup>For details, see [https://www.upress.umn.edu/test-division/mpq/copy\\_of\\_mpq-BF-overview](https://www.upress.umn.edu/test-division/mpq/copy_of_mpq-BF-overview).

*Family Background.* Three indicators of family background assessed at intake were analyzed here. First, Parent Occupational Status was based on mothers' and fathers' reports and coded using the Hollingshead scale (Hollingshead (1957)). We inverted the 1-7 point Hollingshead scale, so that higher scores represented higher occupational status. Individuals were coded as missing if they did not work full-time, were disabled or institutionalized, or reported their occupation as homemaker. The occupation status of the home was taken as the maximum of the two parent reports. Parent College was the number of parents having completed a four-year college degree. Finally, Family Income was measured on a 13-point, self-report scale that ranged from 1 = less than \$10,000 to 13 = Over \$80,000.<sup>12</sup> Information on the income of the twins was collected at the age 29 assessment, and was the answer to the question: "What is your annual income before taxes (in thousands of dollars?)". In the analysis the data on income are translated into dollar amount, then log transformed, and standardized.

#### 4. *PGS* AND EDUCATIONAL ACHIEVEMENT

The GWAS analysis in Lee et al. (2018) was conducted to identify SNP's significantly associated with educational achievement. In this section we test whether the *PGS* constructed using coefficients estimated in that study are useful to predict the same measures of educational achievement (like Education Years and College) as well a related measure (the GPA score) in a sample of subjects that had been excluded from a new meta-analysis using the same methods as in Lee et al. (2018).

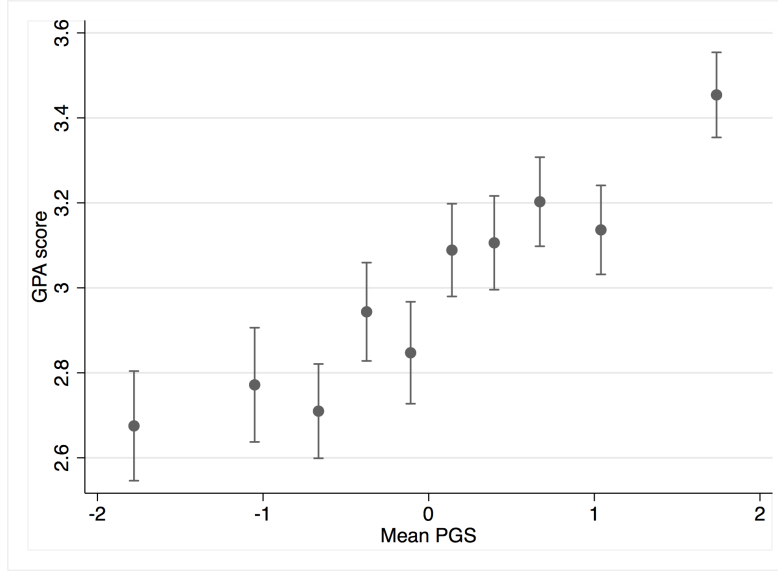
**4.1. GPA score.** A first measure of educational achievement is a summary of the grades in school, collected at age 17 for the twins, the *GPA*-like score, from now *GPA* score, described in detail in section 3.2. We use here the information on grades provided by the mother, as more reliable than the one provided by the child (the correlation between the two measures is 0.76). The score ranges between 0 and 4, mode = 3, mean = 2.984 SD = 0.88. Most of the observations (94.8 per cent) are concentrated in the top three scores (2 or larger). Figure 3 displays mean and 95 % confidence interval for each decile; each group contains between 251 and 253 twins. Note that in this and in the other figures in this section the horizontal distance between two dots is proportional to the difference in *PGS* standardized score.

The difference between the bottom decile (mean = 2.67) and the top (mean = 3.45) is large, considering that the scores are concentrated in the top three. The simple linear regression has an intercept of 2.99 and a slope of 0.23 ( $t = 12.22$ ,  $p < 0.001$ ).

---

<sup>12</sup>The bands were: less than \$10K, \$10,001 to \$15K, \$15,001 to \$20K, \$20,001 to \$25K, \$25,001K to \$30K, \$30,001K to \$35K, \$35,001K to \$40K, \$40,001K to \$45K, \$45,001K to \$50K, \$50,001K to \$60K, \$60,001K to \$70K, \$70,001K to \$80K, more than \$80K.

FIGURE 3. **GPA score and PGS.** *PGS* is standardized with mean 0 and SD 1. The GPA score is in original units, ranging between 0 to 4.

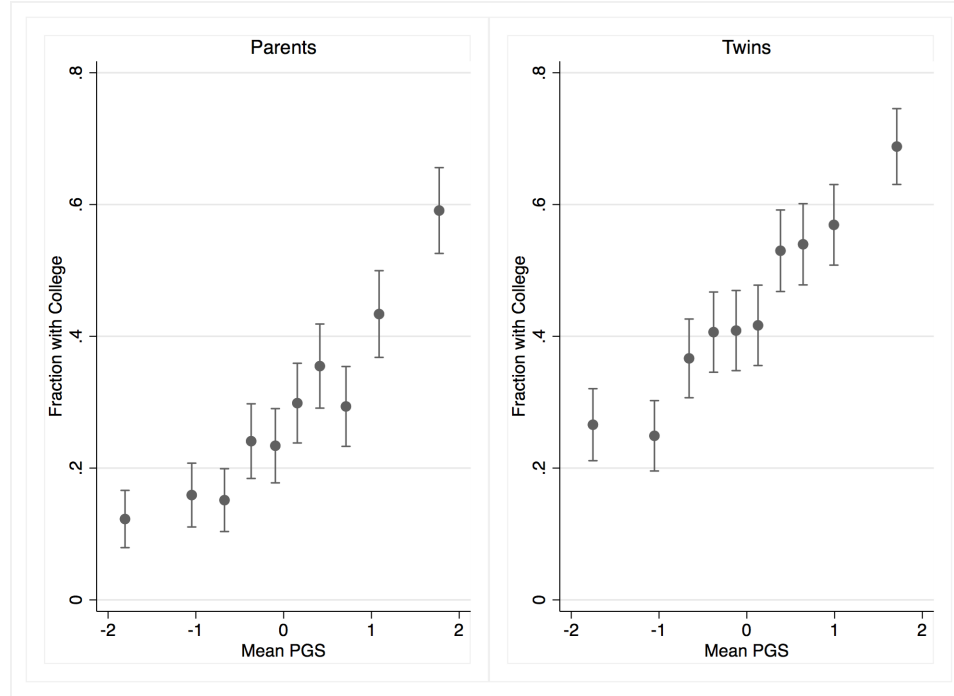


**4.2. College degree.** An important component of educational achievement after school is the achievement of a college degree. We have data on parents' college degree and *PGS* in addition to those of the twins, so we can compare the change over two successive generations of the relation between the two variables. Figure 4 reports the fraction of subjects achieving college for each *PGS* decile, separately for parents and twins, for both sexes. The number of subjects in each group for parents is 220; for Twins is between 251 and 253.

For parents, the fraction reaching a college degree in the lowest decile is 0.12 ( $SE=0.022$ ); in the top decile the fraction is .59 ( $SE=0.033$ ). The corresponding figures for twins are 0.265 ( $SE=0.027$ ) and .688 ( $SE=0.0293$ ), an increase of 14.5 per cent in the lowest and 9.8 per cent in the top. Overall, the curve linking college to *PGS* shifts upward in a parallel way by about 10 to 15 percentage points. In a simple linear regression model of college on the *PGS* the intercept is 28 per cent for parents and 44 per cent for twins ( $SE=0.009$  for both group). The slope is 12.7 per cent for both parents and 13 per cent for twins ( $SE=0.009$  for both group); the interaction term between *PGS* and the indicator variable for parents is not significant. Thus, achieving college is easier (higher intercept) for the younger generation, but the easier access has not made the role of genetic factors, however they may operate, weaker (same slope).

This parallel shift is an average of two quite different changes occurring for the two sexes over the same period. The results for each sex are reported

FIGURE 4. **College degree and  $PGS$ .**  $PGS$  is standardized with mean 0 and SD 1. “Fraction with college” is mean college achievement for each group.

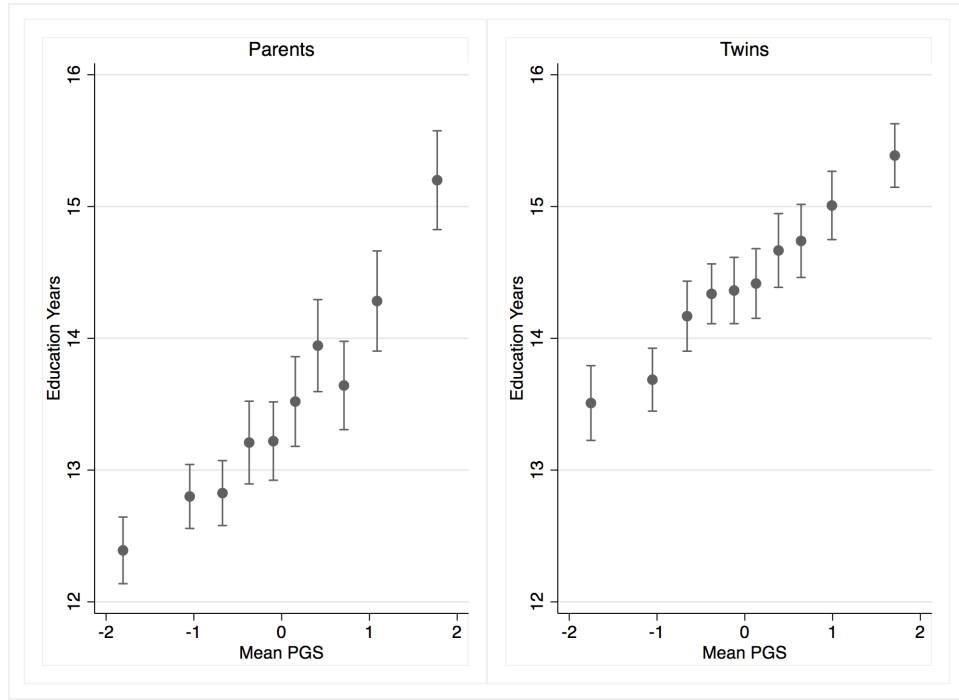


in the Appendix, section A.2. The conclusion of the analysis can be summarized in the observation that the slope of college on  $PGS$  is 0.12 ( $SE=0.014$ ) for fathers and 0.10 ( $SE=0.014$ ) for sons; it is 0.094 ( $SE=0.012$ ) for mothers and .11 ( $SE=0.013$ ) for daughters. Overall, a stable relationship of approximately 10 per cent, with a slight decline for sons and a slight increase for daughters (for the latter, the intercept shifted to 0.509 from the 0.255 for mothers). In conclusion, the larger college graduation rate for women has been the result of a more favorable relation with  $PGS$  over the two generations. The analysis is made more precise by the findings in Tables 2 and 3 (see in particular models (6) in both tables).

Finally, we consider a robustness check: the details and the relation between college and  $PGS$  may depend on the somewhat arbitrary way in which we decide to split the sample of subjects into deciles (rather than, say, quintiles). For comparison, the lowess is reported in the Appendix, figure 11. The conclusions we reached with the analysis based on deciles are confirmed; the effect size is similar, and the curves for the twins are parallel shifts of those for the parents.

**4.3. Education Years.** If we consider education years, we find a relation with *PGS* broadly similar to the one we have seen for college, as figure 5 shows. Slightly different is the evolution over time of the relation for parents and twins between years of education and *PGS*. The slope of the years of education on *PGS* is 0.75 for parents and 0.55 for twins (so the curve is substantially flatter for twins, with interaction term  $-18.6$  per cent ( $p = 0.005$ ); the intercept is similar (13.5 for parents and 14.4 for twins).

FIGURE 5. **Education Years and *PGS*.** The variable Education Years is defined in table 1.



## 5. IDENTIFYING THE PATH FROM *PGS* TO EDUCATION

We have seen estimates indicating a substantial and significant relation between the *PGS* score summarizing information on the genotype, filtered though the coefficients derived from the education *GWAS* on three measures of educational achievement. Since the *GWAS*-derived weights are the same for individuals in our sample, when we try to predict the difference in educational achievement of two individuals based on the *PGS* we are only using information on the two individuals' genotype. In this sense, the analysis in the previous section identifies purely genetic effects.



While these results clearly indicate a link between genotype and education, they are completely silent on the pathway through which a potential effect operates. However, identifying these paths is a crucial pre-condition to any attempt to derive policy conclusions from the findings. From our previous analysis no firm conclusion can be derived, precisely because we have yet no information on the way the effect operates. To illustrate, one may consider two completely unrelated, plausible but both hypothetical pathways.

In an extreme case, differences in educational achievement due to genetic factors are only due to discrimination, which is based on genetic differences (or else they would not be visible through the lens of the *PGS*), which however has no bearing on the ability to reach education. Factors like color of the eyes, of the hairs and of the skin, or height are in large part genetic in origin, but have no impact on the ability or inclination to achieve education. If the pathway from genotype to education was entirely of this type, then a clear implication for policy is that equality of educational outcome can be achieved at no cost. At the other extreme we can consider a case where differences in educational achievement due to genetic factors are produced by differences in individual characteristics that are essential prerequisites for educational achievement. Intelligence, and features of personality such as Conscientiousness or Constraint are examples of such characteristics. This scenario does not preclude of course policies aimed at reducing inequality of educational outcomes. However, and this is very different from the previous scenarios, the policies to be adopted would be potentially costly.

In this section we identify how much of the effect of *PGS* on educational achievement can be attributed to factors such as Personality traits and Intelligence, as well as family environment variables like parents' education and family income.

**5.1. From Intelligence and Personality to education.** Tables 2 and 3 below report the regression analysis for *GPA* score and College for twins; 4 reports the analysis of College achievement for parents. In the Appendix, section A.4, Table 17 reports the ordered logit analysis for Education Years and Table 18 the corresponding analysis for parents. The analysis of these three distinct measures of educational achievement supports specific common conclusions.

First, Intelligence and Personality traits explain a substantial part of the outcome. The coefficient of *IQ* score is significant in the full model in all cases, including the parents. The coefficients of Positive Affect, Negative Affect and Constraint are significant in all cases in the model where only Intelligence and Personality are introduced (model 3 for both twins and parents), in particular when variables describing attitude to academic work (Academic Effort and Academic Problems) and socialization problems

(Externalizing) are excluded. These latter variables are closer to the educational outcome, hence they naturally capture part of the explanatory power of Personality and Intelligence.

Second, only part of the explanatory power of the *PGS* is contained in the variables Intelligence and Personality. The coefficient of *PGS* remains significant and large in all cases, including parents, after these controls are introduced, although of course introducing them reduces the coefficient of the genetic information in *PGS*.

Third, the sizes of the effects of Intelligence and Personality are comparable. In the regression for *GPA* score of twins (table 2) the coefficient of *IQ* score is around 28 per cent (model (2) and (4)); the sum of PA and CN is 27.9 per cent (model (3)), and NA has a significant coefficient of 10.5 (with negative sign in the original variable). In the table 3 for College of twins the odds ratio for *IQ* is between 3.18 (model (2)) and 25.1 (model 4), whereas the product of PA and CN is 3.02, and NA (reversed) has a coefficient 1.69.

Fourth, family environment matters even after we control for the information in genetic data given by *PGS*, Intelligence and Personality. Two variables describe family environment in our data, and the possibly independent and specific contribution to success in education of the twins: family income and education of parents. Both variables have considerable and significant explanatory power even after we condition on *PGS*, as well as Intelligence and Personality traits (with the exception of the role of family income in the *GPA* score).

Fifth, for measures of educational achievement, and for both parents and twins, the information contained in the *PGS* has explanatory power even after we condition on Personality, Intelligence and Family Income in the full model (see in particular model 5 for twins in tables 2 and 3, and model 3 in table 4).

Finally, the role of Intelligence and Personality is qualitatively consistent for twins and parents, although the size of the coefficients changes. There are some striking and interesting exceptions, most notable the role of *MPQ CN* (Constraint) for twins and parents (see in particular model (3) of Table 3 compared to model (2) of Table 4). Decomposing the effect of Constraint in the three components (see section 3.3) of Control, Traditionalism and Harm Avoidance shows that its negative effect on college achievement in parents is the result of a positive effect of Control (as expected, and consistent with the finding for twins) and a negative effect (resulting from cultural factors) of Traditionalism and a negative effect of Harm Avoidance (as expected).

**5.2. From *PGS* to personality.** In the previous section we have examined the link between Personality traits and Intelligence on the one hand, and educational outcome on the other. We now consider the other natural link: that between *PGS* and these two sets of individual characteristics. Figure 6 shows the scatter-plot for the *IQ* score, separately for parents and twins. The correlation is around 30 per cent. It is interesting to note that the

correlation is very similar for both groups; we may contrast this with the difference in the relation between educational outcomes and *PGS* that we have seen in section 4, that showed instead significant differences.

Figure 7 shows the corresponding scatter-plot for the average of the soft skill indexes (these data are only available for twins). The correlation is between one third and one half the one we have found for *IQ*.

Table 5 summarizes the analysis of the effect of *PGS* on different traits. It presents the coefficient of univariate regressions of *PGS* on each of the indicated traits. By far the largest coefficient is the one for *IQ*. Those for the three MPQ meta-traits are all in the expected direction, but smaller in size; the one for Constraint, surprisingly is small and not significant. In line

**TABLE 2. GPA score for Twins. Panel on PGS, IQ and Personality.** Education of parents is the average of years of education of the parents. All variables, including College of parents, are standardized to mean zero and SD 1. The signs of MPQ NA, Externalizing and Academic problems are reversed.

	(1)	(2)	(3)	(4)	(5)	(6)
	b/se	b/se	b/se	b/se	b/se	b/se
PGS	0.247*** (0.025)	0.173*** (0.025)	0.241*** (0.024)	0.164*** (0.023)	0.098*** (0.019)	0.067*** (0.025)
IQ		0.282*** (0.023)		0.293*** (0.022)	0.175*** (0.018)	0.192*** (0.018)
MPQ PA			0.063*** (0.019)	0.049*** (0.018)	0.007 (0.015)	0.013 (0.015)
MPQ NA			0.105*** (0.019)	0.086*** (0.018)	0.026* (0.015)	0.024 (0.015)
MPQ CN			0.216*** (0.021)	0.243*** (0.020)	0.045** (0.018)	0.033* (0.018)
Externalizing					0.041* (0.024)	0.034 (0.023)
Academic effort					0.494*** (0.023)	0.475*** (0.023)
Academic problems					0.116*** (0.021)	0.122*** (0.021)
Education of parents					0.052** (0.021)	0.053** (0.021)
Family Income					0.037 (0.023)	0.028 (0.023)
Male						-0.248*** (0.041)
Male $\times$ PGS						0.067* (0.036)
Constant	0.041 (0.028)	0.052* (0.027)	0.149*** (0.028)	0.166*** (0.027)	0.016 (0.025)	0.132*** (0.031)
N	1801	1801	1801	1801	1801	1801

with what we have seen in figure 7, the coefficient for the Soft Skill index is significant and in size about one third of that for *IQ*.

**5.3. Mediation Analysis.** *Mediation analysis* tests the hypothesis that the way in which an independent variable (*IV*) affects a dependent variable (*DV*) may depend on the intervention of an intermediate mediating variable (*MV*).<sup>13</sup> A mediating variable should affect the value of the dependent

<sup>13</sup>A different approach to mediation analysis is *Causal* mediation analysis: see Hayes (2009) for an exposition. The results in this case confirm those obtained from the standard, Sobel-Goodman, method. They are available from the authors upon request.

**TABLE 3. Probability of College for Twins, Logit regression on PGS, IQ and Personality. Odds ratios reported.** Standard error of OR in parenthesis. Education of parents is the average of years of education of the parents. All independent variables, including College of parents, are standardized to mean zero and SD 1. The signs of MPQ NA, Externalizing and Academic problems are reversed.

	(1) b/se	(2) b/se	(3) b/se	(4) b/se	(5) b/se	(6) b/se
PGS	2.732*** (0.350)	2.187*** (0.289)	2.758*** (0.360)	1.927*** (0.251)	1.686*** (0.214)	1.269 (0.210)
IQ		3.185*** (0.439)		2.517*** (0.340)	2.173*** (0.288)	2.286*** (0.311)
MPQ PA			1.645*** (0.171)	1.437*** (0.158)	1.429*** (0.156)	1.470*** (0.162)
MPQ NA			1.692*** (0.179)	1.415*** (0.155)	1.361*** (0.147)	1.361*** (0.148)
MPQ CN			1.836*** (0.208)	1.153 (0.151)	1.238 (0.161)	1.200 (0.159)
Externalizing				1.367* (0.225)	1.335* (0.216)	1.320* (0.215)
Academic effort				3.703*** (0.674)	3.285*** (0.583)	3.182*** (0.572)
Academic problems				1.308* (0.193)	1.336** (0.194)	1.342** (0.197)
Education of parents					2.049*** (0.277)	2.057*** (0.280)
Family Income					1.489*** (0.213)	1.479*** (0.214)
Male						0.590** (0.151)
Male × PGS						1.858** (0.456)
Constant	0.841 (0.103)	0.872 (0.110)	1.192 (0.158)	0.754* (0.124)	0.726** (0.118)	0.915 (0.182)
N	1838	1838	1838	1838	1838	1838

variable, but should also be affected by the *IV*. The composition of the effect from *IV* to *MV* with that from *MV* to *DV* is the *indirect effect* of the *IV* on *DV*. In our application, the *IV* is the *PGS*, the dependent variable may be any of the variables of interest in our analysis, such as the *GPA* score, college degree, or the education years. The mediating variables we consider are natural candidates for the role of carrying at least part of the effect from *IV* to *DV*, such as Intelligence and Personality traits. Once we decompose the total effect from *IV* to *DV* into the direct and

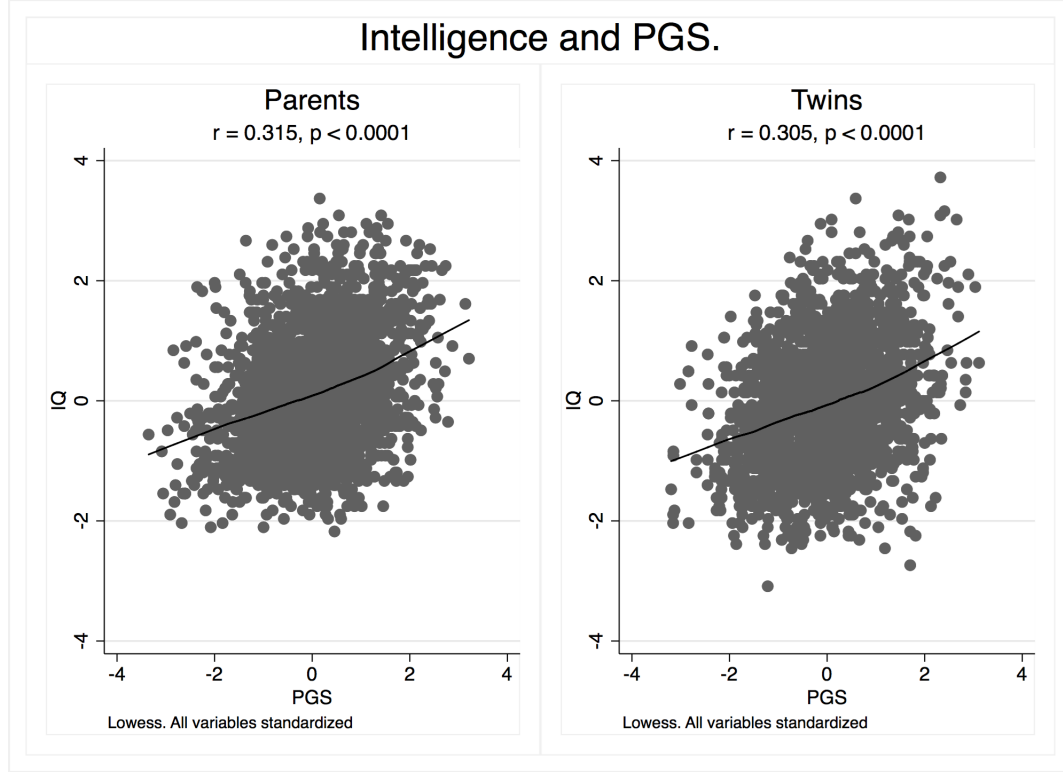
TABLE 4. **Probability of College for Parents, Logit regression on IQ and Personality. Odds ratios reported.** Standard error of OR in parenthesis. All independent variables are standardized to mean zero and SD 1. The sign of MPQ NA is reversed.

	(1)	(2)	(3)	(4)
	b/se	b/se	b/se	b/se
PGS	1.803*** (0.163)	2.337*** (0.209)	1.769*** (0.160)	1.673*** (0.196)
IQ	4.883*** (0.612)		4.640*** (0.585)	4.641*** (0.588)
MPQ PA		1.443*** (0.124)	1.443*** (0.128)	1.445*** (0.129)
MPQ NA		1.286*** (0.115)	1.155 (0.106)	1.155 (0.106)
MPQ CN		0.591*** (0.057)	0.869 (0.085)	0.875 (0.088)
Male				1.024 (0.161)
Male $\times$ PGS				1.138 (0.194)
Constant	0.149*** (0.022)	0.246*** (0.032)	0.159*** (0.025)	0.157*** (0.028)
N	1970	1970	1970	1970

TABLE 5. **Dependent variables: IQ, MPQ, school attitudes. Each variable regressed on PGS. OLS, Twins only** All variables are standardized to mean zero and SD 1.

	IQ	PA	NA	CN	Soft	Ext	AC Eff	Ac Pr
	b/se	b/se	b/se	b/se	b/se	b/se	b/se	b/se
PGS	0.271*** (0.022)	0.065*** (0.024)	-0.046* (0.024)	-0.010 (0.022)	0.093*** (0.027)	-0.065*** (0.023)	0.139*** (0.026)	-0.105*** (0.026)
Const.	-0.041 (0.025)	0.110*** (0.025)	0.274*** (0.025)	-0.408*** (0.023)	-0.001 (0.029)	-0.374*** (0.026)	0.038 (0.029)	-0.030 (0.028)
N	2265	2265	2265	2265	1838	1838	1838	1838

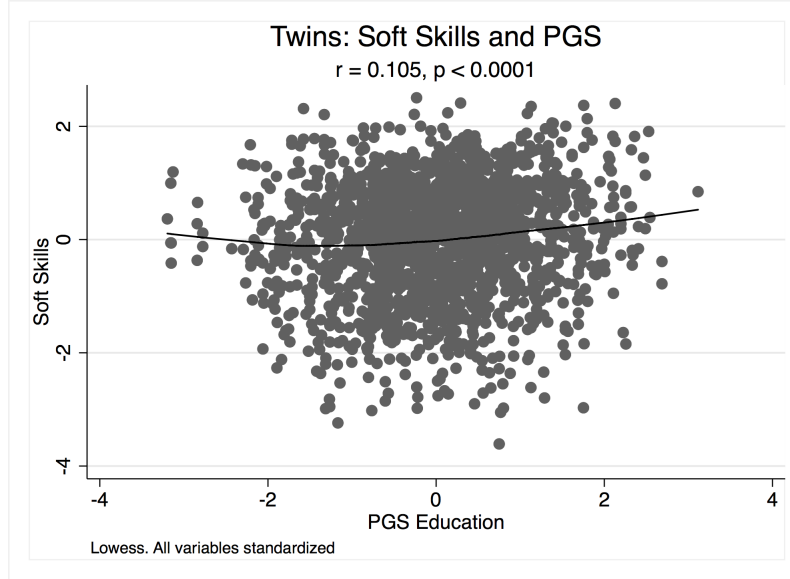
FIGURE 6. **Intelligence and PGS.** Parents (left panel) and twins (right panel) separately.



indirect effect, we have a quantitative estimate of the fraction of the effect of  $IV$  that we can explain (through the action of the  $MV$ ) and the fraction that we can call “direct” and is therefore no further explained. In our data we have detailed information on these variables. *Multiple* mediation analysis considers several mediating variables simultaneously. The results on multiple mediation analysis are reported in section A.7 of the Appendix.

We consider here the indirect effect of  $IQ$  as a  $MV$  with Education Years as  $DV$ ; it is instructive to compare the effects for parents and for twins. For the entire sample (parents and twins) the coefficient of the regression of  $IQ$  on  $PGS$  is .302 ( $SE = 0.014$ ,  $Z = 21.49$ ). As we know, this coefficient is remarkably stable over time: it is .301 ( $SE = 0.019$ ,  $Z = 15.09$ ) for twins, and .304 ( $SE = 0.019$ ,  $Z = 15.38$ ) for parents. The coefficient of the regression of Education Years on  $IQ$  is for the entire sample .301 ( $SE = 0.014$ ,  $Z = 21.76$ ). This coefficient changes in the two generations. It is .187 ( $SE = 0.018$ ,  $Z = 15.03$ ) for twins, but is substantially larger for parents: .477 ( $SE = 0.019$ ,  $Z = 15.38$ ). The proportion of the total affect that is

FIGURE 7. **Soft skills and PGS; Twins only.** The information on soft skills is not available for parents.



explained by the mediating variable *IQ* is .34 (Sobel-Goodman test (*SGT*) = 0.09,  $SE = 0.006$ ,  $Z = 15.29$ ), but lower in twins: .24 ( $SGT = 0.05$ ,  $SE = 0.006$ ,  $Z = 8.51$ ), and higher in parents: .47 ( $SGT = 0.14$ ,  $SE = 0.011$ ,  $Z = 12.82$ ). In all cases, bootstrapping confirms the significance of the estimated coefficients.

In conclusion, the role of *IQ* as mediating variable between genetic factors and Education Years is approximately  $\frac{1}{3}$ , but this is the average of almost  $\frac{1}{2}$  for parents and, one generation later,  $\frac{1}{4}$  for twins.

## 6. PASSIVE GENE ENVIRONMENT CORRELATION

An additional insight on the role of genetic factors can be gained by considering the information on the *PGS* of the parents. Clearly, all the information on the genotype of the parents that is relevant for the determination of the genotype of the twins is rendered irrelevant by the direct information that we have on the genotype of the twins. The genotype of the parents determines a probability distribution on the genotype of the offspring (with higher correlation between *MZ* twins, but the same distribution for each individual considered separately); the same holds, in a slightly more complicated way, for *PGS*, which is a linear functions of the genotype. The information on *PGS* of the children gives us a more precise information than the one provided by the *PGS* of the two parents, because it takes into

account the specific realization of the random variable “genotype of the offspring”. However, as we discussed in detail in section 2.8, the genotype of the parents can very well have an additional indirect effect of the phenotype of interest of the off-springs (educational achievement in our case) through the effect of the environment on the phenotype (passive Gene-Environment correlation,  $r_{GE}$ ).<sup>14</sup> We analyze this information in the case of several variables: *GPA* in Table 6, number of education years in Table 7, Intelligence in Table 8, and finally probability of college degree in table 9.

TABLE 6. **GPA on PGS of Twin and PGS of parents, IQ and Soft Skills.** All variables, including GPA, are standardized to mean zero and SD 1.

	(1) b/se	(2) b/se	(3) b/se	(4) b/se	(5) b/se
PGS	0.213*** (0.034)	0.199*** (0.036)	0.130*** (0.031)	0.224*** (0.033)	0.139*** (0.031)
PGS mother	0.064* (0.033)	0.062* (0.035)	0.025 (0.029)	0.011 (0.034)	0.000 (0.029)
PGS father	0.052 (0.035)	0.036 (0.036)	−0.008 (0.030)	−0.014 (0.036)	−0.039 (0.031)
IQ			0.222*** (0.022)		0.207*** (0.023)
Soft Skills Index			0.414*** (0.021)		0.408*** (0.021)
Education of parents				0.160*** (0.033)	0.088*** (0.028)
Family Income				0.095** (0.037)	0.041 (0.031)
Constant	0.038 (0.031)	0.061* (0.032)	0.044* (0.026)	0.012 (0.030)	0.031 (0.026)
N	1583	1393	1393	1583	1393

The estimated coefficients for all variables have a common natural pattern across the models. The *PGS* of each of the parents is significant and in some case not negligible (for example, for education years and *IQ*) when no control is introduced (model 1); therefore there is an effect of the genotype on education that is operating through the environment that parents provide. The last model (6) shows however that when controls are introduced the *PGS* of the parents is small and usually insignificant: hence the controls we have introduced capture most of this environmental effect. The difference between unconditional and conditional model persists even if we use a restricted sample in the latter model (model 2); this model is introduced to make the comparison between the first and the last model more

<sup>14</sup>The role of family environment is considered in detail in the companion paper Willoughby et al. (2018).)



transparent, since in model (2) and (5) the sample size is the same. The control variables *IQ* and the soft-skill index (model 3) are significant but do not reduce substantially the estimated coefficients of the parents. Instead, as natural, the controls variables parents' education and family income do reduce the size and significance of the coefficients of the parents' *PGS*, as the passive *rGE* model predicts. It is interesting to note also that the *PGS* of the mother and the father have distinct effects.

**TABLE 7. Education Years on PGS of Twin and PGS of parents, IQ and Soft Skills.** All variables, including Education Years, are standardized to mean zero and SD 1.

	(1)	(2)	(3)	(4)	(5)
	b/se	b/se	b/se	b/se	b/se
PGS	0.106*** (0.031)	0.122*** (0.034)	0.082** (0.032)	0.121*** (0.030)	0.097*** (0.031)
PGS mother	0.106*** (0.027)	0.084*** (0.029)	0.059** (0.027)	0.045* (0.026)	0.025 (0.027)
PGS father	0.091*** (0.028)	0.054* (0.030)	0.024 (0.028)	0.009 (0.028)	-0.023 (0.028)
IQ			0.139*** (0.023)		0.112*** (0.023)
Soft Skills Index			0.230*** (0.021)		0.218*** (0.021)
Education of parents				0.182*** (0.025)	0.114*** (0.025)
Family Income				0.116*** (0.027)	0.088*** (0.028)
Constant	0.298*** (0.023)	0.327*** (0.025)	0.317*** (0.023)	0.271*** (0.023)	0.296*** (0.023)
N	1690	1337	1337	1690	1337

For college, to compare coefficients we report in Table 9 the result for the linear model on the standardized (mean zero and SD 1) college variable. The logit analysis is reported in Table 20 of the Appendix.

## 7. FIXED EFFECTS ANALYSIS IN *DZ* TWINS

*DZ* twins offer a uniquely informative way for the analysis of the effect of genetic variables on educational achievement. *DZ* twins share many significant variables: date and condition of birth, family background and very similar family environment in the following years. Therefore, a fixed effect analysis of measures of educational achievements regressed on *PGS*, once restricted to *DZ* twins, will control for the effect of environmental factors common to the two twins. The correlation of *PGS* between twins (once we restrict to *DZ* twins) is high, but there is sufficient variability to allow robust analysis. Figure 8, top panel, illustrates.

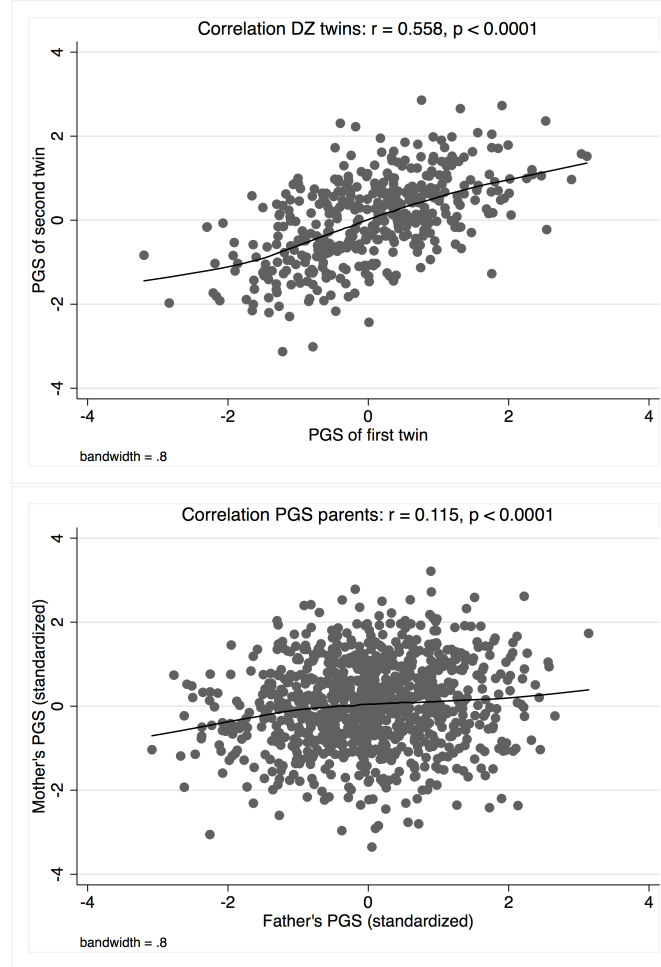
TABLE 8. **IQ on PGS of Twin and PGS of parents, IQ and Soft Skills.** All variables are standardized to mean zero and SD 1.

	(1)	(2)	(3)	(4)	(5)
	b/se	b/se	b/se	b/se	b/se
PGS	0.186*** (0.033)	0.178*** (0.037)	0.173*** (0.037)	0.203*** (0.032)	0.187*** (0.036)
PGS mother	0.068** (0.032)	0.059* (0.035)	0.053 (0.035)	0.000 (0.032)	−0.000 (0.035)
PGS father	0.115*** (0.032)	0.122*** (0.036)	0.117*** (0.036)	0.032 (0.033)	0.046 (0.037)
Soft Skills Index			0.087*** (0.024)		0.074*** (0.024)
Education of parents				0.242*** (0.031)	0.221*** (0.034)
Family Income				0.022 (0.033)	0.005 (0.038)
Constant	−0.015 (0.029)	−0.009 (0.032)	−0.011 (0.031)	−0.033 (0.028)	−0.029 (0.031)
N	1809	1415	1415	1809	1415

TABLE 9. **College on PGS of Twin and PGS of parents, IQ and Soft Skills. Linear Model** All independent variables are standardized to mean zero and SD 1.

	(1)	(2)	(3)	(4)	(5)
	b/se	b/se	b/se	b/se	b/se
PGS	0.082*** (0.018)	0.089*** (0.020)	0.061*** (0.019)	0.093*** (0.017)	0.071*** (0.018)
PGS mother	0.055*** (0.016)	0.046*** (0.018)	0.031* (0.016)	0.014 (0.015)	0.007 (0.016)
PGS father	0.041** (0.016)	0.021 (0.018)	0.002 (0.017)	−0.010 (0.016)	−0.029* (0.017)
IQ			0.099*** (0.013)		0.081*** (0.013)
Soft Skills Index			0.141*** (0.013)		0.135*** (0.012)
Education of parents				0.124*** (0.014)	0.088*** (0.015)
Family Income				0.057*** (0.016)	0.039** (0.017)
Constant	0.485*** (0.014)	0.504*** (0.015)	0.500*** (0.014)	0.470*** (0.013)	0.487*** (0.014)
N	1809	1415	1415	1809	1415

FIGURE 8. **Top Panel:** *PGS* of twin and co-twin, DZ twins only. **Bottom Panel:** *PGS* of mother and father. The continuous line is for both panels the lowess at bandwidth 0.8.



*Genetic Assortative Matching.* We have seen in section 2.7 the theoretical estimate of the correlation among *DZ* twins. The difference in *PGS* correlation reported in figure 8 and the predicted correlation with random assortative matching (which is  $\frac{1}{2}$ ) is 0.0575 and it must be due to the assortative matching among parents. In our case we are considering not the genome-wide correlation<sup>15</sup> but the one between *PGS* of parents. The bottom panel of Figure 8 illustrates the correlation, and reports the correlation coefficient between *PGS* of the two parents ( $r = 0.115$ ). As discussed recently in the

<sup>15</sup>See Robinson et al. (2017), Supplementary Note, page 12.

literature (see Abdellaoui et al. (2014), Robinson et al. (2017)), the estimate of genetic assortative mating can be influenced by population stratification, which may produce spurious correlation. For example the genetic assortative mating estimated in Domingue et al. (2014) becomes insignificant when a control with principal components is performed <sup>16</sup>. In section A.6 of the appendix we report the controls for PC in our data; Table 21 shows that the estimated correlation in *PGS* of spouses is robust to such control. This correlation is to be expected, given the strong correlation between education years of the two parents, reported in the bottom panel of Figure 15 in the Appendix, section A.5. For education years, the correlation coefficient is  $r = 0.504$ .

*GPA in DZ twins.* The overall (that is, in the sample including *MZ* and *DZ*) correlation for *GPA* score between twin and co-twin is .67 ( $p < 0.0001$ ), with a value .785 ( $p < 0.0001$ ) for *MZ* and .492 ( $p < 0.0001$ ) for *DZ*. The fixed effects panel regression of the *GPA* score is reported in Table 10. The coefficient of *PGS* is large (27.9 per cent) and significant ( $p < 0.0001$ ) in the model with *PGS* as only explanatory variable. The coefficient remains substantial and significant even after we condition on the additional information on the subject provided by *IQ* scores and personality measures. The coefficient of the *IQ* score is considerably larger than those of the other traits.

*IQ in DZ twins.* For *IQ*, the overall (in the sample including *MZ* and *DZ*) correlation between twin and co-twin is .701 ( $p < 0.0001$ ), with a value of .785 ( $p < 0.0001$ ) for *MZ* and .552 ( $p < 0.0001$ ) for *DZ*. The scatter-plot and lowess for *IQ* of twin and co-twin is presented in Figure 9. Table 11 reports the fixed effects regression for standardized *IQ* score. The coefficient of *PGS* is substantial (13.8 per cent) and significant ( $p = .005$ ) in the model with the single *PGS* variable. The other models are reported for comparison with different explanatory variables.

*Within and Between effects in DZ twins.* In the analysis with fixed effects we are forced to ignore variables that are identical for the two twins, such as family income and education of parents. In Tables 12 and 13 we report the results of the regression in the *DZ* sample of standardized *GPA* and *IQ* on the difference from the twin average, where we add (see models 2 and 3) the twin average value of *PGS*. The use of this method for the analysis of twin data is presented in is discussed in Gurrin et al. (2006) (see also Carlin et al. (2005) for the reasons to recommend the use of a general model that includes separate regression coefficients for within-twin-pair and between-pair effects). The table confirms the result of the fixed effects regression (the coefficients for the difference between the own *PGS* and the average

---

<sup>16</sup>See Section S2 Principal Components of Domingue et al. (2014), Table S1. These are the same tests we use in Table 21.

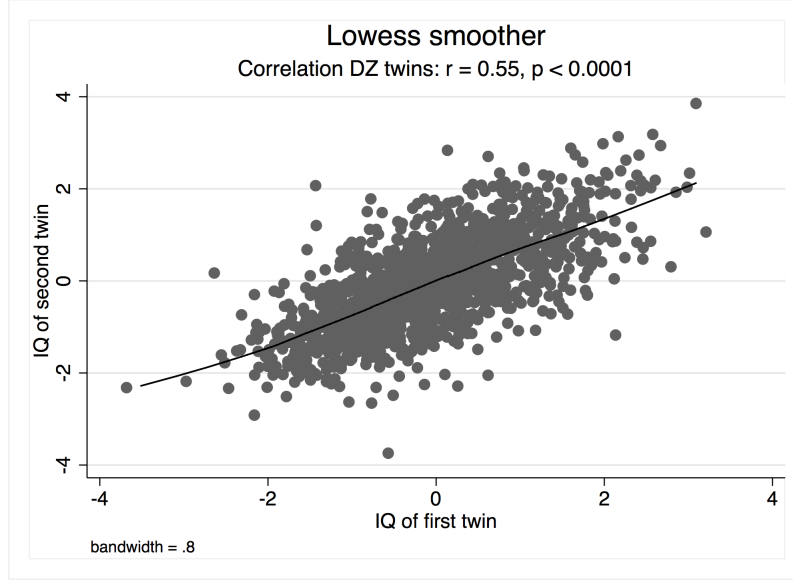
TABLE 10. **GPA score: Fixed effects analysis in DZ twins.** All variables are standardized to mean zero and SD 1.

	(1)	(2)	(3)
	b/se	b/se	b/se
PGS	0.279*** (0.054)	0.191*** (0.053)	0.142*** (0.043)
IQ		0.334*** (0.053)	0.188*** (0.044)
MPQ PA		0.075* (0.042)	0.059* (0.034)
MPQ NA		0.004 (0.044)	-0.051 (0.036)
MPQ CN		0.210*** (0.045)	0.005 (0.042)
Externalizing at 17			0.105* (0.061)
Academic effort at 17			0.469*** (0.057)
Academic problems at 17			0.100** (0.047)
Constant	-0.026 (0.027)	0.103*** (0.034)	-0.042 (0.041)
N	682	630	590

TABLE 11. **IQ score: Fixed effects analysis in DZ twins.** All variables are standardized to mean zero and SD 1.

	(1)	(2)	(3)
	b/se	b/se	b/se
PGS	0.138*** (0.048)	0.146*** (0.053)	0.097* (0.058)
MPQ PA		0.043 (0.043)	0.024 (0.047)
MPQ NA		0.105** (0.042)	0.108** (0.049)
MPQ CN		-0.087** (0.044)	-0.161*** (0.057)
Externalizing at 17			-0.153* (0.082)
Academic effort at 17			0.266*** (0.075)
Academic problems at 17			0.095 (0.064)
Constant	-0.067*** (0.024)	-0.054 (0.034)	-0.048 (0.056)
N	802	723	601

FIGURE 9. **IQ score of twin and co-twin, DZ twins only.** The continuous line is the lowess at bandwidth 0.8.



value is similar to the one estimated in the panel fixed effects),<sup>17</sup> but adds a precise estimate of the role of the mean *PGS* for the twin pair, as well as an estimate of the role of family income and parents' education. The effect of the twin mean *PGS* is larger than that of the difference for *GPA* and *IQ*. The effect of parents' education is larger than that of family income for both *GPA* and *IQ*, although the first is significant in both cases, and not negligible for *GPA*.

Table 19 in the Appendix (section A.4) reports the results of the regression of *IQ* on one twin's *PGS* and the *PGS* of the co-twin. The regression is equivalent to the one reported in table 13, but it provides a different way to evaluate the effect of the own *PGS* score for a twin (coefficient own *PGS* is 0.267,  $p < 0.0001$ , coefficient of other twin's *PGS*: 0.108,  $p = 0.006$ ).

*College in DZ twins.* In Table 14 we report the conditional logit analysis of the college variable. Odds ratios are reported; the odds ratio for *PGS* is substantial and significant, and remains so after we condition on information on *IQ* and personality traits. Note that in the case of college degree also

<sup>17</sup>Note that the estimated coefficients for model 1 in table 10 and the one in model 1 in 12, which are remarkably similar, do not need to be the same, because the independent variable is different in the two cases, because the fixed effect estimator corrects for the overall mean of the dependent variable.

*IQ* has a significant effect, and substantially larger than that of the other traits.

In Table 15 we report the logit analysis of the fixed effects considering as regressors the difference from the average *PGS* and the average *PGS*, and then controlling for family income and parents' education. Again, odds ratios are reported; they are not negligible, and significantly different from 1. Although a direct comparison with the coefficients estimated for *GPA* and *IQ* is not possible, a regression with a linear model shows that the coefficient of the income is 7 per cent, that of the parents' education 14 per cent; the coefficient of the *PGS* difference is 4 per cent, and that of the average *PGS* 13 per cent.

TABLE 12. **GPA and PGS (difference from mean and mean), OLS, DZ twins.** All variables standardized to mean zero and SD 1.

	(1)	(2)	(3)
	b/se	b/se	b/se
Twins' Difference PGS	0.273*** (0.078)	0.276*** (0.074)	0.273*** (0.072)
Twins' Average PGS		0.358*** (0.040)	0.282*** (0.042)
Family income			0.125*** (0.039)
Education of parents			0.145*** (0.039)
Constant	-0.007 (0.038)	-0.032 (0.036)	-0.041 (0.036)
N	687	687	687

TABLE 13. **IQ and PGS (difference from mean and mean), OLS, DZ twins.** All variables standardized to mean zero and SD 1.

	(1)	(2)	(3)
	b/se	b/se	b/se
Twins' Difference PGS	0.138* (0.072)	0.138** (0.068)	0.138** (0.066)
Twins' Average PGS		0.357*** (0.036)	0.268*** (0.038)
Family income			0.059* (0.034)
Education of parents			0.209*** (0.036)
N	802	803	803

TABLE 14. **College and PGS in DZ twins: Conditional logit analysis in DZ twins, odds ratios reported.** All variables standardized to mean zero and SD 1.

	(1) b/se	(2) b/se	(3) b/se
PGS	1.628*** (0.301)	1.508* (0.319)	1.666* (0.439)
IQ		3.179*** (0.982)	3.295*** (1.306)
MPQ PA		1.425 (0.320)	1.581 (0.458)
MPQ NA		1.190 (0.262)	1.406 (0.393)
MPQ CN		1.545** (0.328)	0.999 (0.331)
Externalizing at 17			2.955* (1.827)
Academic effort at 17			1.200 (0.619)
Academic problems at 17			1.225 (0.426)
N	224	192	162

TABLE 15. **College and PGS in DZ twins: Logit analysis in DZ twins, odds ratios reported.** For PGS, both the mean value, and difference from the mean for each twin, is considered. All variables standardized to mean zero and SD 1.

	(1) b/se	(2) b/se	(3) b/se
Twins' Difference PGS	1.188** (0.088)	1.224** (0.098)	1.263*** (0.109)
Twins' Average PGS		2.539*** (0.243)	2.032*** (0.209)
Family income			1.620*** (0.160)
Education of parents			1.941*** (0.174)
Constant	0.792*** (0.056)	0.731*** (0.056)	0.722*** (0.060)
N	814	814	804

## 8. CONCLUSIONS

We have examined the predictive power of the *PGS* for educational achievement, a score constructed for each individual using only the individual's genotype.



We setup the analysis in a natural extension of theories of parental investment and intergenerational mobility (as in Becker and Tomes (1979) and in the literature building on that model), but replaced the *ad hoc* assumption of an asexual  $AR(1)$  process with a fully specified formulation of genetic transmission of skills from a pair of parents in a stable mating equilibrium with preferences consistent with the parental investment model.

Our model provides the basis for an economic analysis of genetic factors in education and intergenerational mobility; it has the virtue of being more realistic than the existing models, and it can now be tested in the data. The predictions of our model of intergenerational mobility differ substantially from the standard model. Most notably, there is no constant heritability coefficient as in the standard model; instead heritability is determined endogenously and depends on the probability distribution of the genotype and on the features of the assortative mating, hence of the mating preferences of the agents. We have concluded in our analysis that the standard model is likely to underestimate the intergenerational elasticity of income. Our model also allows a precise test of important features affecting intergenerational mobility, such as assortative mating and passive gene-environment correlation, which is the effect of genes of parents operating (over and above the direct effect on genes) through the environment provided to children.

Genetic factors measured by the *PGS* have a large effect on educational achievement, for example raising the fraction achieving college from about 20 per cent in the low decile to about 60 per cent in the top decile. The effect has changed in our data in the two generations covered by our data (spanning a period in the USA going from the 70's to the late 90's), mostly producing a shift upward of the curve rather than a change in slope; thus the relative impact of genetic factors across individuals has not decreased, particularly for college degree. These changes have been different in the two sexes, and largely in favor of females.

Very different pathways of the effect of *PGS* could be consistent with this finding: for example, the effect might be entirely due to discrimination operating on individual characteristics that are genetically based but irrelevant for the technology of educational achievement. These discrimination effects are less likely for components that operate through Intelligence and Personality; any fraction of the explanatory power of the *PGS* that can be attributed to the mediation on these individual characteristics is less likely to operate through discrimination. Regression and mediation analysis show that the pathways occur in a significant part through Intelligence and Personality, and that the size of the effect of Intelligence is overall stronger.

Our data include information on the genetic profile of the parents, so we can test directly size and significance of passive gene-environment correlation. This correlation is evident for all the variables considered (*GPA*, education years, college degree but also Intelligence); but the correlation is particularly strong for variables (such as education years) for which parental resources in addition to the genetic endowment are strong. Interestingly, the

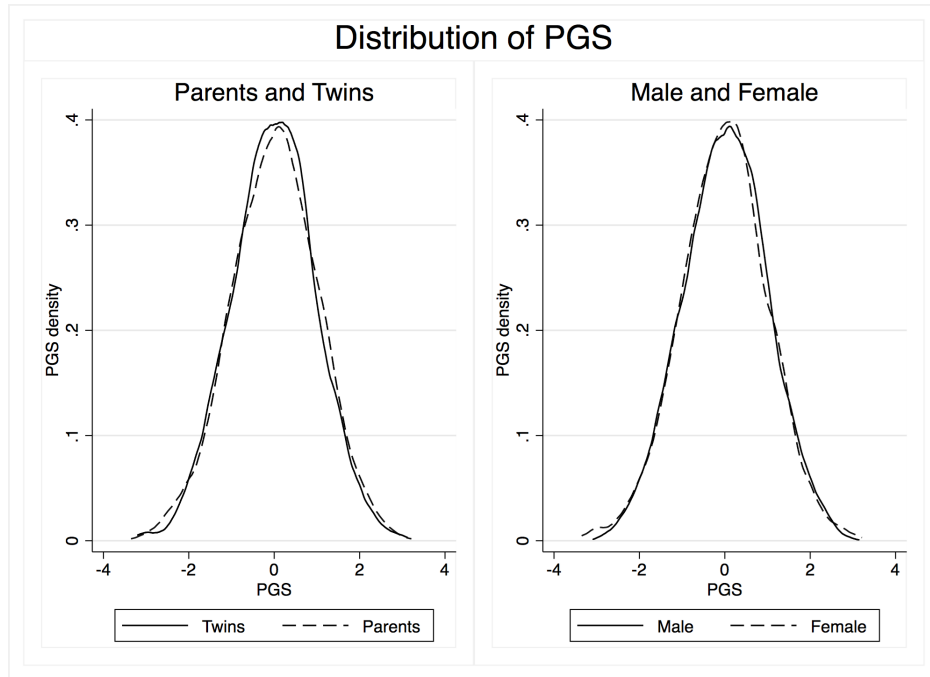
contribution of the genetic endowment of the mother and the father is specific, even for characteristics (such as *IQ*) where the difference might not be expected.

Fixed effects analysis on *DZ* twins shows significant effect of *PGS* on a measure of academic performance at school (the *GPA* score), Intelligence as well as in educational achievement, in particular college degree. This final result provides an important support for our conclusion, since *DZ* twins share very similar environments in their formative years, but are significantly different in genotype, in spite of assortative mating. Our analysis shows that in all cases the additional effect of parents' education and family income are strong and significant, and that of parents' education larger than that of income.

## APPENDIX A. APPENDIX (NOT MEANT FOR PUBLICATION)

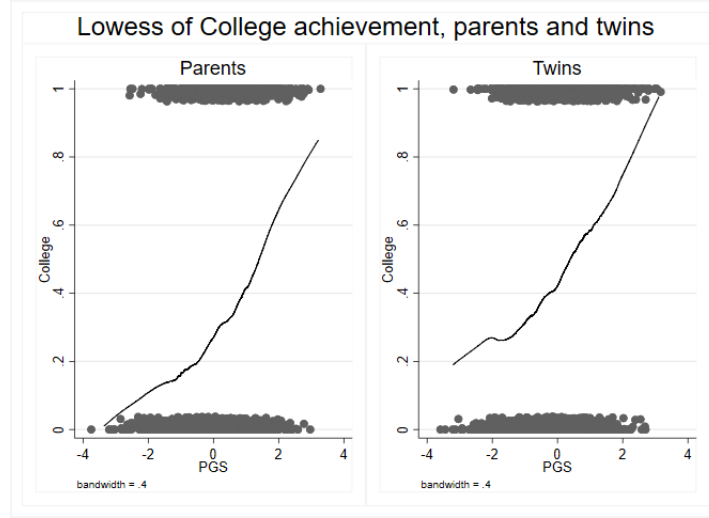
A.1. **Distribution of PGS.** Figure 10 reports the kernel density for parents and twins, male and female separately. The two pairs of distributions are, as natural, very similar.

FIGURE 10. **Distribution of PGS.** Left panel: parents and twins. Right panel: male and female.



**A.2. College achievement and PGS.** The relation between college achievement and PGS displayed in figure 4 of the main text is also robust to different techniques of identifying the relation. Below we display the results of the lowess.

FIGURE 11. **College achievement and PGS.** Top panel: Lowess. Bottom Panel: Fractional Polynomial. The data points in the scatter-plot of the lowess figure are jittered to give some information on the number of subjects with and without college degree, for different values of PGS.

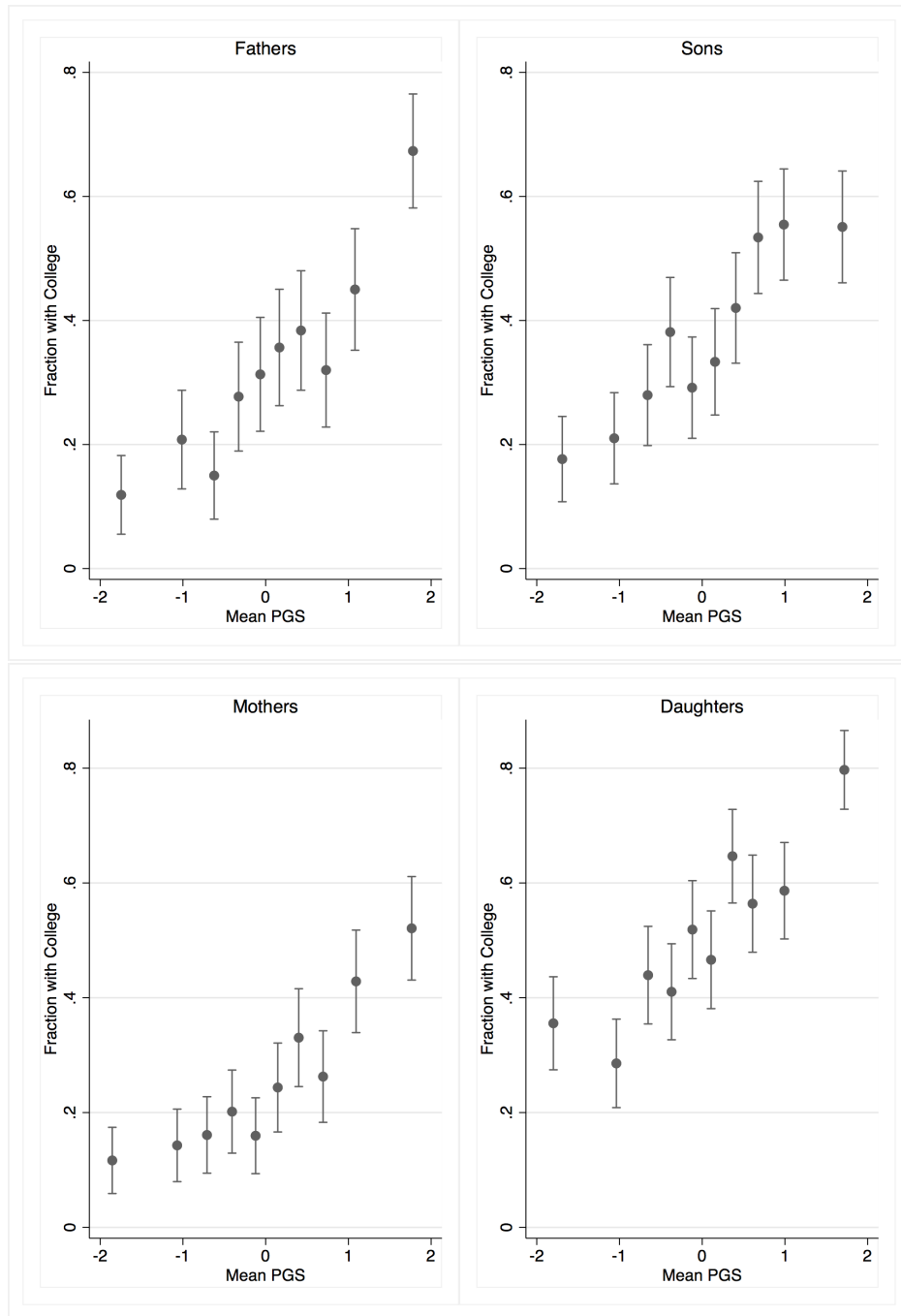


**A.2.1. Differential effect on sexes.** The relation between achievement of college education changed between the two generations (parents and twins), but also it did so in a different way for the two sexes. Results are displayed in Figure 12. The number of subjects in each group for Fathers is 101; for Twins is 118. The corresponding figures are for mothers is 119 and for daughters 134.

For fathers, the fraction reaching a college degree in the lowest decile is .12 (SE .032); in the top decile the fraction is .67 (SE 0.047). The corresponding figures for sons are .18 (SE .035) and .55 (SE .045), an increase of only 5.8 per cent in the lowest and a *decrease* of 12.3 per cent in the top decile.

For mothers, the fraction reaching a college degree in the lowest decile is .116 (SE .029); in the top decile the fraction is .521 (SE 0.046). The corresponding figures for daughters are .355 (SE .041) and .79 (SE .035), an increase of 23.9 per cent in the lowest and an increase of 27.5 per cent in the top decile.

FIGURE 12. **College degree and PGS.** Top Panel, Fathers and Sons. Bottom Panel, Mothers and Daughters.



A.3. **Evidence of  $rGE$ .** Figure 13 reports the scatterplot and lowess of household income and parental  $PGS$ .

FIGURE 13. **Correlation  $PGS$  of parents and family income.** *Household income* is the standardized value of log of the household income in USD.

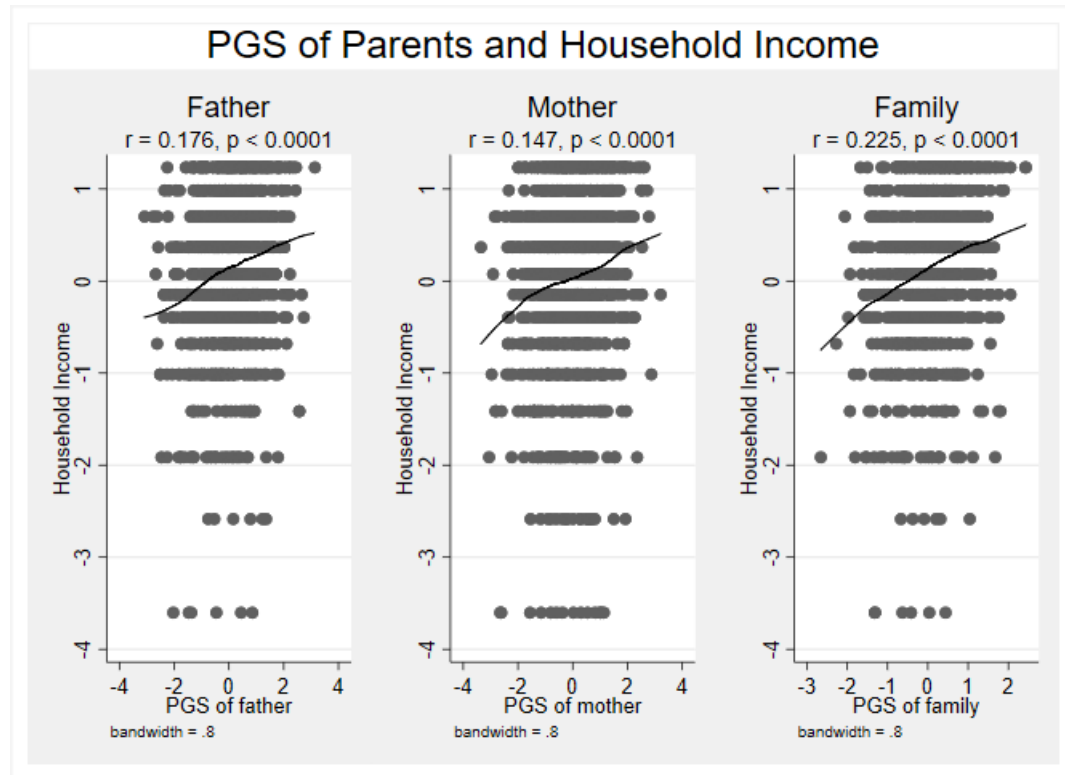


TABLE 16. **Family income and parental GPS.** OLS clustered by family.  
All variables are standardized to mean zero and SD 1.

	(1)	(2)	(3)	(4)
	b/se	b/se	b/se	b/se
PGS mother	0.132*** (0.028)		0.120*** (0.027)	0.062** (0.027)
PGS father		0.151*** (0.029)	0.140*** (0.029)	0.068** (0.030)
Education Parents				0.247*** (0.029)
Constant	0.029 (0.028)	0.127*** (0.027)	0.125*** (0.028)	0.109*** (0.027)
N	2358	2010	1896	1896

## A.4. Regression Analysis.

TABLE 17. **Class of Education of Twins. Ordered logit regression on Personality, IQ and PGS. Odds ratios displayed. SE of OR in parenthesis.** Class 1 (basis): Less than HS, GED, HS; Class 2: HS and Vocational, Community College. Class 3: College and Professional degree. All independent variables are standardized to mean zero and SD 1. To ease comparisons, we changed the sign of the variables MPQ NA.

	(1) b/se	(2) b/se	(3) b/se	(4) b/se	(5) b/se	(6) b/se
ClassEducation						
PGS Education	1.655*** (0.082)	1.485*** (0.076)	1.674*** (0.084)	1.496*** (0.077)	1.414*** (0.075)	1.329*** (0.071)
IQ		1.587*** (0.081)		1.626*** (0.085)	1.492*** (0.081)	1.382*** (0.077)
MPQ PA			1.260*** (0.060)	1.223*** (0.059)	1.150*** (0.058)	1.136** (0.058)
MPQ NA			1.292*** (0.063)	1.248*** (0.061)	1.149*** (0.058)	1.115** (0.057)
MPQ CN			1.432*** (0.075)	1.537*** (0.082)	1.077 (0.067)	1.139** (0.072)
Externalizing					1.286*** (0.092)	1.268*** (0.091)
Academic effort					1.857*** (0.130)	1.748*** (0.123)
Academic problems					1.091 (0.069)	1.099 (0.071)
Education of parents						1.423*** (0.078)
Family Income						1.247*** (0.073)
N	1713	1713	1713	1713	1713	1713



TABLE 18. **Class of Education of Parents. Ordered logit regression on Personality, IQ and PGS. Odds ratios displayed. SE of OR in parenthesis.** Class 1 (basis): Less than HS, GED, HS; Class 2: HS and Vocational, Community College. Class 3: College and Professional degree. All independent variables are standardized to mean zero and SD 1. The sign of the variable MPQ NA is reversed.

	(1)	(2)	(3)
	b/se	b/se	b/se
PGS Education	1.580*** (0.082)	1.890*** (0.094)	1.554*** (0.081)
IQ	2.948*** (0.167)		2.862*** (0.167)
MPQ PA		1.288*** (0.064)	1.301*** (0.068)
MPQ NA		1.278*** (0.067)	1.185*** (0.065)
MPQ CN		0.703*** (0.039)	0.896* (0.053)
N	1970	1970	1970

TABLE 19. **IQ and PGS (own and other twin), all Twins.** All independent variables standardized to mean zero and SD 1.

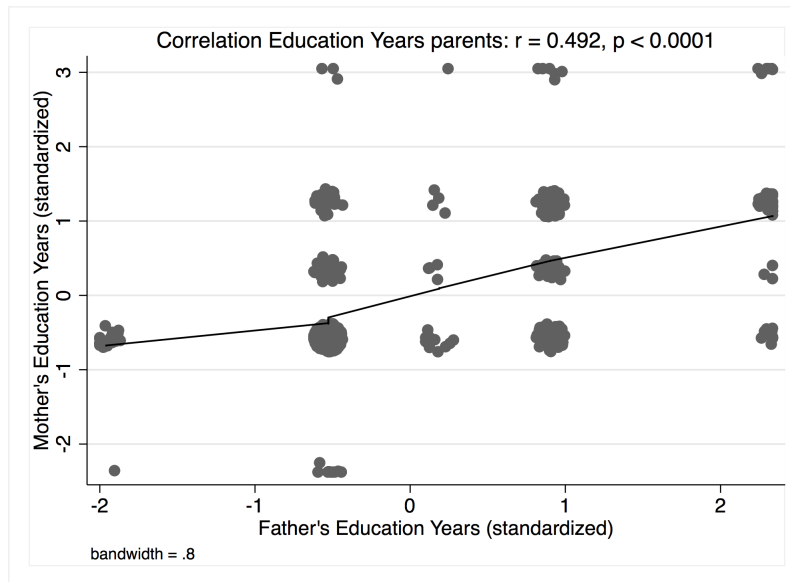
	(1)	(2)	(3)
	b/se	b/se	b/se
Own PRS	0.306*** (0.019)	0.229*** (0.036)	0.194*** (0.035)
Other Twin's PRS		0.091** (0.036)	0.057 (0.035)
Family income			0.045** (0.020)
Education of parents			0.223*** (0.020)
Constant	-0.055*** (0.019)	-0.055*** (0.019)	-0.060*** (0.018)
N	2427	2427	2427

TABLE 20. **College on PGS of Twin and PGS of parents, IQ and Soft Skills. Odds ratios displayed.** All independent variables are standardized to mean zero and SD 1.

	(1)	(2)	(3)	(4)	(5)	(6)
	b/se	b/se	b/se	b/se	b/se	b/se
college						
PGS	1.956*** (0.298)	2.065*** (0.358)	1.745*** (0.308)	2.146*** (0.318)	1.912*** (0.334)	1.801*** (0.317)
PGS mother	1.587*** (0.213)	1.494*** (0.228)	1.328* (0.201)	1.116 (0.143)	1.044 (0.156)	1.239 (0.186)
PGS father	1.427*** (0.196)	1.215 (0.190)	1.045 (0.164)	0.944 (0.126)	0.794 (0.126)	0.973 (0.152)
IQ			2.492*** (0.356)		2.126*** (0.296)	2.414*** (0.341)
Soft Skills Index			3.690*** (0.542)		3.459*** (0.497)	3.540*** (0.515)
Education of parents				2.629*** (0.341)	2.171*** (0.326)	
Family Income				1.665*** (0.226)	1.477** (0.238)	1.860*** (0.304)
Constant	0.881 (0.101)	1.030 (0.135)	0.998 (0.131)	0.791** (0.087)	0.899 (0.116)	0.903 (0.120)
N	1809	1415	1415	1809	1415	1415

A.5. **Assortative mating in Education.** Figure 14 documents the correlation between education years of the two parents in the sample.

FIGURE 14. **Standardized Education Years of parents.** Points in the scatter plot are jittered to indicate the number of observations. The continuous line is the lowess at bandwidth 0.8.



**A.6. Evidence of Genetic Assortative Mating.** Figure 15 displays the scatterplot and lowess of the *PGS* of father and mother.

FIGURE 15. **Correlation of *PGS* of parents.** The continuous line is the lowess at bandwidth 0.8.

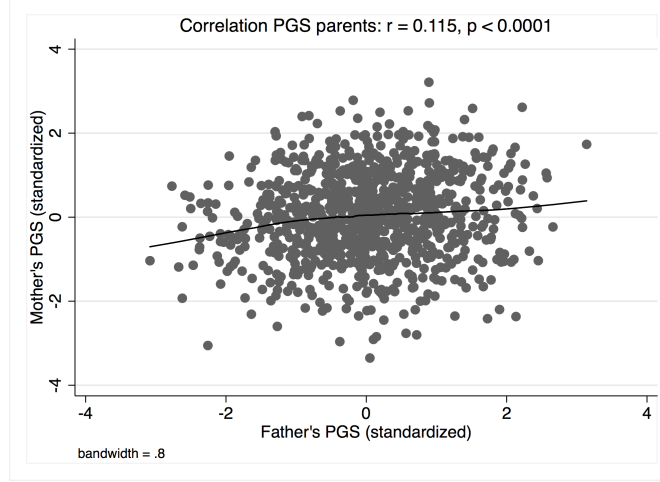


Table 21 shows that the size of the genetic assortative mating is robust to control for possible population stratification. There are 10 principal components; only the first one has significant effects in the regression of the *PGS* of the mother over the *PGS* of the father.

TABLE 21. **PGS of mother on PGS of father controlling for PC distance, all PC's.** OLS . Model 1: no control. Model 2: controlling for interaction of PGS mother and father. Model 3: controlling for square of the difference. Model 4: controlling for absolute value of the difference. Only the first PC (PC1) is reported; all others have insignificant coefficients in all models. All variables are standardized to mean zero and SD 1.

	(1)	(2)	(3)	(4)
	b/se	b/se	b/se	b/se
PGS father	0.118*** (0.033)	0.114*** (0.033)	0.117*** (0.033)	0.117*** (0.033)
Interaction PC1		170.441** (68.664)		
Square Diff. PC1			38.976 (39.652)	
Abs. Val. Diff. PC1				1.748 (2.572)
N	953	953	953	953

**A.7. Mediation Analysis.** We report the details of the estimation of the indirect effects of PGS through Intelligence and Soft skills on GPA and Education Years. The estimates are controlled for family income.

Computing the SE through the delta method (as suggested in Sobel. (1982), Sobel. (1986)) requires assumption of normality of the total and indirect affects, an assumption which is unlikely to hold. To avoid this assumption we bootstrap the sampling distribution as suggested in Shrout and Bolger (2002). Tables 22 and 23 reports the results for the three different methods suggested in Efron (1987), Efron and Tibshirani (1993).

*A.7.1. Mediation Analysis of GPA.*

**TABLE 22. Bootstrap of Indirect coefficients of PGS on GPA score.** (P): percentile confidence interval; (BC): bias-corrected confidence interval; (BCa) bias-corrected and accelerated confidence interval.

	Coefficient	Bias	Bootstrap SE	95% CI		
Indirect IQ	.065	-.000019	.0072	.0513	.0799	(P)
				.049	.0805	(BC)
				.0495	.0804	(BCa)
Indirect Soft	.0424	-.00012	.0114	.0197	.0654	(P)
				.0196	.0654	(BC)
				.0197	.0731	(BCa)
Tot Indirect	.107	-.000202	.0135	.0804	.134	(P)
				.0802	.134	(BC)
				.0802	.134	(BCa)

*A.7.2. Mediation Analysis of Education Years.*

**TABLE 23. Bootstrap of Indirect coefficients of PGS on Education Years.** (P): percentile confidence interval; (BC): bias-corrected confidence interval; (BCa) bias-corrected and accelerated confidence interval.

	Coefficient	Bias	Bootstrap SE	95% CI		
Indirect IQ	.0362	.000036	.0059	.0252	.0489	(P)
				.0256	.0495	(BC)
				.0256	.0495	(BCa)
Indirect Soft	.0231	0.00003534	.0067	.0102	.044	(P)
				.0101	.0365	(BC)
				.0100	.0364	(BCa)
Tot Indirect	.0594	0.00007	.00913	.0420	.0777	(P)
				.0419	.0776	(BC)
				.0419	.0777	(BCa)

## REFERENCES

- ABDELLAOUI, A., K. J. H. VERWEIJ, AND B. P. ZIETSCH (2014): “No evidence for genetic assortative mating beyond that due to population stratification,” *Proceedings of the National Academy of Sciences*, 111, E4137–E4137.
- AIYAGARI, S. R., J. GREENWOOD, AND N. GUNER (2000): “On the State of the Union,” *Journal of Political Economy*, 108, 213–244.
- BECKER, G. AND N. TOMES (1979): “An Equilibrium Theory of the Distribution of Income and Intergenerational Mobility,” *Journal of Political Economy*, 87, 1153–89.
- (1986): “Human Capital and the Rise and Fall of Families,” *Journal of Labor Economics*, 43, S–1–39.
- BECKER, G. S. (1973): “A theory of marriage: part I,” *Child Development Perspectives*, 81, 813–846.
- BERG, J. J., A. HARPAK, N. SINNOTT-ARMSTRONG, A. M. JOERGENSEN, H. MOSTAFAVI, Y. FIELD, E. A. BOYLE, X. ZHANG, F. RACIMO, J. K. PRITCHARD, AND G. COOP (2018): “Reduced signal for polygenic adaptation of height in UK Biobank,” *bioRxiv*.
- BLACK, S. E. AND P. J. DEVEREUX (2011): “Chapter 16 - Recent Developments in Intergenerational Mobility,” in *Handbook of Labor Economics*, ed. by D. Card and O. Ashenfelter, Elsevier, vol. 4, Part B, 1487 – 1541.
- CARLIN, J. B., L. C. GURRIN, J. A. STERNE, R. MORLEY, AND T. DWYER (2005): “Regression models for twin studies: a critical review,” *International Journal of Epidemiology*, 34, 1089–1099.
- CESARINI, D. AND P. M. VISSCHER (2017): “Genetics and educational attainment,” *npj Science of Learning*, 2, 1–7.
- CRONBACH, L. (1951): “Coefficient alpha and the internal structure of tests,” *Psychometrika*, 16, 297–334.
- CROW, J. F. AND M. KIMURA (1970): *An introduction to Population Genetics Theory*, Harper and Row.
- DISNEY, E. R., I. J. ELKINS, M. MCGUE, AND W. G. IACONO (1999): “Effects of ADHD, conduct disorder, and gender on substance use and abuse in adolescence,” *American Journal of Psychiatry*, 156, 1515–1521.
- DOMINGUE, B. W., J. FLETCHER, D. CONLEY, AND J. D. BOARDMAN (2014): “Genetic and educational assortative mating among US adults,” *Proceedings of the National Academy of Sciences*, 111, 7996–8000.
- DUDBRIDGE, F. (2013): “Power and Predictive Accuracy of Polygenic Risk Scores,” *PLoS Genet*, 9, 1–17.
- EFRON, B. (1987): “Better bootstrap confidence intervals,” *Journal of the American Statistical Association*, 82, 171–185.
- EFRON, B. AND R. TIBSHIRANI (1993): *An introduction to the bootstrap*, Chapman and Hall.
- FERNANDEZ, R., N. G. GUNER, AND J. KNOWLES (2005): “Love And Money: A Theoretical And Empirical Analysis Of Household Sorting And

- Inequality," *The Quarterly Journal of Economics*, 120, 273–344.
- FERNANDEZ, R. AND R. ROGERSON (2001): "Sorting and Long-Run Inequality," *The Quarterly Journal of Economics*, 116, 1305–1341.
- GREENWOOD, J., N. GUNER, AND J. A. KNOWLES (2003): "More on Marriage, Fertility, and the Distribution of Income," *International Economic Review*, 44, 827–862.
- GREENWOOD, J., N. GUNER, G. KOCHARKOV, AND C. SANTOS (2016): "Technology and the Changing Family: A Unified Model of Marriage, Divorce, Educational Attainment, and Married Female Labor-Force Participation," *American Economic Journal: Macroeconomics*, 8, 1–41.
- GURRIN, L. C., J. B. CARLIN, J. A. C. STERNE, G. S. DITE, AND J. L. HOPPER (2006): "Using Bivariate Models to Understand between- and within-Cluster Regression Coefficients, with Application to Twin Data," *Biometrics*, 62, 745–751.
- HAYES, A. F. (2009): "Beyond Baron and Kenny: Statistical Mediation Analysis in the New Millennium," *Communication Monographs*, 76, 408–420.
- HECKMAN, J. J. AND T. KAUTZ (2012): "Hard evidence on soft skills," *Labour Economics*, 19, 451 – 464.
- HOLLINGSHEAD, A. (1957): *Two Factor Index of Social Position*, Hollingshead.
- IACONO, WILLIAM ND CARLSON, S. R., J. TAYLOR, I. J. ELKINS, AND M. MCGUE (1999): "Behavioral disinhibition and the development of substance use disorders: Findings from the Minnesota Twin Family Study." *Development and Psychopathology*, 11, 869–900.
- JAFFEE, S. AND T. PRICE (2007): "Geneenvironment correlations: a review of the evidence and implications for prevention of mental illness," *Molecular Psychiatry*, 12, 432–442.
- JOHNSON, W., M. M. MCGUE, AND W. G. IACONO (2004): "Genetic and environmental influences on academic achievement trajectories during adolescence," *Developmental Psychology*, 42.
- LEE, J. ET AL. (2018): "Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals," *Nature Genetics*.
- LOURY, G. (1981): "Intergenerational Transfers and the Distribution of Earnings," *Econometrica*, 49, 843–67.
- MCGUE, M., D. IRONS, AND W. IACONO (2014): "The adolescent origins of substance use disorders: A behavioral genetic perspective." in *Genes and the motivation to use substances*, ed. by S. F. Stoltenberg, New York: Springer.
- MULLIGAN, C. B. (1997): *Parental Priorities and Economic Inequality*, University of Chicago Press.
- (1999): "Galton versus the Human Capital Approach to Inheritance," *Journal of Political Economy*, 107, S184–S224.

- NAGYLAKI, T. (1992): *Introduction to Theoretical Population Genetics*, Springer Verlag.
- OKBAY, A. ET AL. (2016): “Genome-wide association study identifies 74 loci associated with educational attainment,” *Nature*.
- PLOMIN, R., J. DEFRIES, AND J. C. LOEHLIN (1977): “Genotype-environment interaction and correlation in the analysis of human behavior,” *Psychological Bulletin*, 84, 309–322.
- REICH, W. (2000): “Diagnostic Interview for Children and Adolescents (DICA),” *Journal of the American Academy of Child and Adolescent Psychiatry*, 39, 59 – 66.
- RIETVELD, C. ET AL. (2013): “Individuals Identifies Genetic Variants Associated with Educational Attainment,” *Science*, 340, 1467–1471.
- ROBINSON, M. ET AL. (2017): “Genetic evidence of assortative mating in humans,” *Nature Human Behaviour*, 1.
- SATTLER, J. M. (1974): *Assessment of children’s intelligence*, Saunders.
- SCARR, S. AND K. MCCARTNEY (1983): “How people make their own environments: a theory of genotype greater than environment effects,” *Child Development*, 54, 424–435.
- SHROUT, P. AND N. BOLGER (2002): “Mediation in experimental and non-experimental studies: New procedures and recommendations,” *Psychological Methods*, 7, 422–445.
- SOBEL, M. E. (1982): “Asymptotic confidence intervals for indirect effects in structural equations models,” in *Sociological Methodology*, ed. by S. Leinhardt., San Francisco: Jossey-Bass.
- (1986): “Some new results on indirect effects and their standard errors in covariance structure model,” in *Sociological Methodology*, ed. by N. Tuma, Washington, DC: American Sociological Association.
- SOHAIL, M., R. M. MAIER, A. GANNA, A. BLOEMENDAL, A. R. MARTIN, M. C. TURCHIN, C. W. K. CHIANG, J. N. HIRSCHHORN, M. DALY, N. PATTERSON, B. NEALE, I. MATHIESON, D. REICH, AND S. R. SUNYAEV (2018): “Signals of polygenic adaptation on height have been overestimated due to uncorrected population structure in genome-wide association studies,” *bioRxiv*.
- SOLON, G. (1992): “Intergenerational Income Mobility in the United States,” *American Economic Review*, 82, 393–408.
- (2004): “A model of intergenerational mobility variation over time and place,” in *Generational income mobility in North America and Europe*, ed. by M. Corak, Cambridge: Cambridge University Press.
- SPITZER, R., J. WILLIAMS, M. GIBBON, AND M. FIRST (1992): “The structured clinical interview for dsm-iii-r (scid): I: history, rationale, and description,” *Archives of General Psychiatry*, 49, 624–629.
- TELLEGEN, A. AND N. G. WALLER (2008): “Exploring personality through test construction: Development of the Multidimensional Personality Questionnaire,” in *The SAGE Handbook of Personality Theory and Assessment. Volume 2: Personality Measurement and Testing*, ed. by G. J.



- G. J. Boyle, G. Matthews, and D. Saklofske, London: Sage.
- VILHJLMSSON, B. J. ET AL. (2015): “Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores,” *The American Journal of Human Genetics*, 97, 576–592.
- WELNER, Z., W. REICH, B. HERJANIC, K. JUNG, AND H. AMADO (1987): “Reliability, validity, and parent-child agreement studies of the Diagnostic Interview for Children and Adolescents (DICA).” *Journal of the American Academy of Child and Adolescent Psychiatry*, 26, 649–653.
- WILLOUGHBY, E. A., M. MCGUE, W. G. IACONO, A. RUSTICHINI, AND J. J. LEE (2018): “The role of parental genotype in predicting offspring years of education: Evidence for passive geneenvironment correlation,” Tech. rep., University of Minneapolis, Department of Psychology, Minneapolis.

(Aldo Rustichini) DEPARTMENT OF ECONOMICS, UNIVERSITY OF MINNESOTA, 1925 4TH STREET SOUTH 4-101, HANSON HALL, HANSON HALL, MINNEAPOLIS, MN, 55455  
*E-mail address:* aldo.rustichini@gmail.com

(William G. Iacono) DEPARTMENT OF PSYCHOLOGY, UNIVERSITY OF MINNESOTA, 75 EAST RIVER RD., MINNEAPOLIS, MN 55455  
*E-mail address:* wiacono@umn.edu

(James Lee) DEPARTMENT OF PSYCHOLOGY, UNIVERSITY OF MINNESOTA, 75 EAST RIVER RD., MINNEAPOLIS, MN 55455  
*E-mail address:* leex2293@umn.edu

(Matt McGue) DEPARTMENT OF PSYCHOLOGY, UNIVERSITY OF MINNESOTA, 75 EAST RIVER RD., MINNEAPOLIS, MN 55455  
*E-mail address:* mcgue001@umn.edu