EDUCATIONAL ATTAINMENT AND INTERGENERATIONAL MOBILITY: A POLYGENIC SCORE ANALYSIS

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

Abstract. We extend a standard model of parental investment and intergenerational mobility to include a fully specified genetic analysis of skill transmission. The model's predictions differ substantially from standard model's. The coefficient of intergenerational income elasticity (IGE) may be larger than in the standard model, and depends on distribution of the genotype. The distribution of genetic endowments may be stratified according to income. The model is tested on data, including genetic information, of twins and their parents, estimating how IGE is affected by genetic factors, and how environment and genes interact. The effect of intelligence is substantially stronger than other traits'.

(Aldo Rustichini) Department of Economics, University of Minnesota, 1925 4th Street South 4-101, Hanson Hall, Hanson Hall, Minneapolis, MN, 55455
Email address: aldo.rustichini@gmail.com

(William G. Iacono) Department of Psychology, University of Minnesota, 75 East River Rd., Minneapolis, MN 55455
Email address: wiacono@umn.edu

(James J. Lee) Department of Psychology, University of Minnesota, 75 East River Rd., Minneapolis, MN 55455
Email address: leex2293@umn.edu

(Matt McGue) Department of Psychology, University of Minnesota, 75 East River Rd., Minneapolis, MN 55455
Email address: mcgue001@umn.edu

Date: February 19, 2023.

We thank three anonymous referees and the editor who gave advice that lead to a complete reorganization of the paper, of the analysis as well as the exposition. We are in particular thankful to one of the referees that urged us to clarify the relationship between the research reported here and earlier publications. We thank Aysu Okbay for generously running the meta-analysis on the data, Peter Visscher for a clarification on Robinson et al. (2017), Philippe Köllinger for the help in the process. We also thank Tom Holmes, Andrea Ichino, Chris Phelan, Joel Waldfogel, 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 ($DA05147$).

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 [1] 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 which are not significant by $GWAS$ standards.

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).

Theoretical Framework. We set up the investigation in a fully specified model of parental investment in education of children. Some classical papers establishing this tradition are Becker and Tomes (1979), Loury (1981), Becker and Tomes (1986). Important developments of the early model are, among many, in Solon (1992), Mulligan (1997), Mulligan (1999), Solon (2004), Black and Devereux (2011), Black et al. (2017)). Our model differs from the existing ones in the field in two respects, both introduced because we 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 in the the model a theory of mating [2] (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),

[1] The threshold is 5×10^{-8}; the factor 10^{-8} corrects (Bonferroni) for multiple comparisons.

[2] In this paper two terms, matching and mating are used interchangeably, as synonymous for partnership among parents. The reason for the coexistence of the two terms is that the "matching" is used more frequently in the economics literature, and "mating" in behavioral genetics. In each instance we use the term most appropriate in the context.

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)). From our vantage point, after so much research, we can revisit the classical debate between Goldberger (1989) and Becker (1989), and realize that both models were, in some important measure, imprecise. We take this opportunity to illustrate the implications of our work. [3] Becker had in mind the auto-regressive process assumed in his earlier work (Becker and Tomes (1979), Becker and Tomes (1986)), that we discuss more in detail later (section 3.3). In his thought provoking 1985 Woytinsky lecture, Goldberger suggests a modification of Galton (1886) Regression Towards Mediocrity, [4] presenting the Galton's argument that the characteristic of the individual is some weighted average of the characteristics of the entire history of ancestors. But the expectation of the child's phenotype conditional on the entire history of genotypes of the ancestors is equal to the expectation conditional on the parents' genotype only. It also has a precise form, [5] which is neither the one in Becker and Tomes (1979) nor the one in Goldberger (1989).

Empirical Questions. Within our theoretical framework we address two basic sets of questions. First, how much of the variance in income and educational achievement is explained by the PGS, and how does family structure affect the transmission? Similar questions have been investigated using the same data in McGue et al. (2017) and McGue et al. (2020), but simply examining correlational results, rather than tests of a well specified model of parental investment. Tests of the effectiveness of the polygenic score in predicting a variety of variables including economic success are presented in Rietveld et al. (2013), Okbay et al. (2016), Lee et al. (2018), Kong et al. (2018), Willoughby et al. (2021), wealth in Barth et al. (2020), social mobility of children compared to parents, (Belsky et al. (2018)), as well as health outcomes, such as Barcellos et al. (2018) (earlier contributions on the issue of health conditions and academic performance using genetic markers is in Ding et al. (2009), Fletcher and Lehrer (2011)). 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 modified by assortative mating among parents, and the correlation among their genotypes? And finally, how is the effect of genes mediated by the direct effect

[3]The discussion between Goldberger and Becker centered on two main points: whether adopting a utility maximization framework makes a difference for the predictions of the theory, and what is the stochastic process of skill. For the first, we adopt here the utility maximization setup, but, as long as the comparison of policies is not explicitly modeled, choosing one of the other makes little difference. We focus here on the second point.

[4]Page 505 in Goldberger (1989).

[5]See e.g., equation 10.104 in Nagylaki (1992)

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 in each individual? Recall that the score is built on a simple statistical association between genotype and the phenotype of interest, in our case success in education, and no mechanism underlying the association is identified. 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. We now review our answers to these questions.

Outline of Main Results. We develop (section 2) a model of intergenerational mobility, building on classical parental investment models, but replacing their *ad-hoc* skill transmission equation with a precise and correct model of genetic inheritance from the two parents. In the model the coefficient of persistence of skills is endogenous, depending on the distribution of the genotypes in the populations; thus most of the conclusions of the classical model are now invalid. We provide the correct predictions.

An important component of the theory is the model of assortative matching among parents according to characteristics, some endowed with a natural order (such as income and skills), and some not (such as personality traits, or physical appearance); we show how this affects the distribution of the genotypes at the invariant distribution of the system. A state of the system is described by a joint probability on genotypes and endogenous variables, such as income and education. Because of with assortative matching, the transition function is non linear, so existence of a stationary distribution is not simple. We prove its existence and some basic properties.

At the stationary distribution, within each class of matching, alleles are in Hardy-Weinberg equilibrium. [6] More notably, the frequency of alleles with positive effect on educational attainment, and thus on income, positively correlates with income. The correlation is stronger, the stronger the

[6]The definition of Hardy-Weinberg is recalled later, see section 3.

effect of the allele. These results identify a powerful force producing lasting inequality. which has been ignored so far, and is absent by assumption in standard models. [7]

The model leads to a natural empirical test, using data described in section 9. Information on genotype of individuals is summarized into a polygenic score obtained from a large $GWAS$ on educational attainment. All the predictions of the model are tested in a unique set of data in which we have complete genetic information on parents and children, in addition to information on education, personality traits, intelligence, family environment and income. We estimate (section 5) the intergenerational elasticity coefficient of income, which is in the lower end of existing estimates for the overall USA. We compare it to the effect size of genetic factors measured by the polygenic score; we find that the latter is approximately half of that of income. In section 6 we identify the pathways of the effect on income through human capital formation.

In section 7 use the twin structure of our data to check for the robustness of the results and investigate *passive gene* $\times$ *environment correlation*, that is, how the genetic endowment of the parents affects the phenotype of the children through the family environment. Natural and significant channels of this effect are education of parents and their income, and we prove this channel is significant. However, there is no additional residual channel through family environment in addition to these two. When we study the pathways of the genetic effect measured by the PGS, we find that after correcting for measurement errors, the effect from genotype to educational and economic success is mostly mediated by intelligence, and only weakly by non-cognitive skills. Conclusions are presented in section 8.

2. Genetic Skill Transmission and Parental Investment

We begin with the conceptual and theoretical structure for our empirical analysis, introducing a model and an equilibrium concept. The complete model to be tested is presented in section 2.7. Our first aim is to show that the standard analysis of parental investment in education, and intergenerational mobility (as pioneered in Becker and Tomes (1979), where the skill transmission follows a simple $AR(1)$ process), should be modified –if one wants to avoid significant misunderstandings–to take into a account a fully specified genetic mechanism of skill transmission. A core feature of the model we propose is the combination of the theory of marriage (Becker (1973)) to predict mating, with a model of genetic transmission. A comparison of the prediction of the two models is provided in section 3.3; there we show that they differ substantially on key predictions, for instance on inter-generational mobility.

[7] We use standard model in a broad sense here, which includes Goldberger (1989)'s model.

Our model has several components. After defining the basic environment (section 2.1), in section 2.2 we describe how the skill of the children are affected by genetic endowment inherited by parents, family environment and random events. Then, in section 2.4, we describe the decision of parents to invest into education of the children.

2.1. **Setup.** A population of individuals, constant in number over time, is organized into households. A household maximizes a utility function of own consumption and future income of two children, which in turn is affected by genetic endowment of the children, parental investment in education, 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 has little importance when we study parental investment, [8] but becomes important when we study the correlation of skill and income across siblings. We denote y the natural log of the income (so this value ranges in the real line), E consumption expenditure, I parental investment in education of children and h human capital measured by the education level. ϵ^e and ϵ^y denote the random shocks to education and income respectively: each one is $i.i.d.$ across periods and the two are independent within periods. Subscripted α's denote productivity parameters of the variable in the index; so α_I, α_h denote positive real numbers associated to parental investment and human capital. $\delta \in (0,1)$ is the discount factor. A vector of real numbers $\theta = (\theta^1, \ldots, \theta^{n_1}, \theta^{n_1+1}, \ldots, \theta^n)$ 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 et al. (2013)). Skills enter linearly into the production of the education level though an n-dimensional vector of coefficients α_θ. 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^i_f and mother, y^i_m. [9] The precise form of the combination will be specified later. We denote $\mathbf{E}$ the expectation of a random variable.

We emphasize that the model is not a two-period models, but an overlapping generations model, so each individual appears in the model as a child and then as a parent, and the model describes behavior in both stages of life. So when we model, for instance, how genetic factors affect skill formation, human capital accumulation and income of children, we also model how the same variables have been determined for the parents of these children. We make full use of this in some crucial step; for example in section A.5, where we describe how genetic factors affect behavior both as children and as parents.

[8] Note however that two children who are also twins have the same age, so the parental investment in this case does not concern two individuals of different age, as instead typical for siblings.

[9] The use of letters f and m avoids confusion with the family index.

2.2. **Skill Transmission.** We replace the standard $AR(1)$ mechanism of skill transmission (discussed more extensively in section 3.3 below) with a detailed model where the skill vector θ 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. [10] If K is the number of loci, a genotype is a $g \in G \equiv \{0,1,2\}^K$, so $g = (g(k) : k = 1, \ldots, K)$. Here $0, 1, 2$ refers to the count of one of the alleles in a bi-allelic system (a $GWAS$ typically deals with variants, SNP's, that are bi-allelic in the analysis). The joint distribution of genotypes of the two children, given the genotype of the two parents, depends on the twin type, that may be monozygotic, MZ or dizygotic, DZ. To describe how the distribution is determined we start with the function from parents' genotype to the probability over genotypes of an individual offspring, given by a function H from $G \times G$ to $\Delta(G)$:

$$H : (g_m, g_f) \mapsto H(g_m, g_f). \tag{1}$$

We will write $H(\cdot|g_m, g_f)$ when we want to indicate explicitly the set on which this measure operates. H follows well known rules of Mendelian inheritance (see for example Nagylaki (1992), or Crow and Kimura (1970)); for instance if $K = 1$, so that $G = \{0, 1, 2\}$, then $H(\cdot|1, 1)$ is $(0.25, 0.5, 0.25)$, and so $H(2|1, 1) = 0.25$. Similarly, $H(\cdot|0, 2)$ is $(0, 1, 0)$.

The map in equation (1) is well defined only under the assumption, which we make, that the distribution across loci is independent. Simple examples show that we may have two different haplotype pairs which induce the same genotype profile (g_m, g_f) for the parents but, without this assumption, induce different elements in $\Delta(G)$ for the children.

2.3. **Polygenic Scores.** Let w denote a n-valued function determining skills as function of the genotype g. The polygenic scores are denoted by $w(g)$. They are computed assuming additivity across loci and within each locus, so that:

$$w(g) = \sum_{k=1}^{K} \alpha(k) g(k) \tag{2}$$

[10] In the context of the twins-studies model, the integration of parental investment models with a more realistic model of skill transmission has been explored in Rustichini et al. (2017), where the realistic model of genetic transmission is used to provide a justification for the standard ACE models in twin studies in the context of economic analysis of parental investment. However, in in Rustichini et al. (2017) there is no analysis of the invariant measure produced by assortative mating according to characteristics (which we consider one of the main contributions of this paper), nor a comparison of the predictions of the standard skill transmission model in the tradition of Becker and Tomes (1979). Finally, the empirical analysis does not make any use of the genetic information used in this paper.

where α is a vector of parameters. The values w are latent variables, and they would be of little use if we did not have an estimate. We will rely on estimates, called *estimated polygenic score* of the true effect $w(g)$ of the genotype g, is:

$$PGS(g) = \sum_{k=1}^{K} \beta(k) g(k) \tag{3}$$

where β's are weights derived from a genome-wide association study ($GWAS$). We should note that the weights obtained in a $GWAS$ do not give full account of the variability in educational attainment. There may be rare variants (Yengo et al. (2020)) as well as structural variants (Chiang et al. (2017)) that are not well captured by a $GWAS$ study. [11] We let X_j^i denote vector of variables associated to twin ij . These variables may be observable or not, and may include for instance the parents' education, personality traits of the child, family's social status, and so on. Also let Π a matrix with n rows, F a family specific n-dimensional vector (common to both twins in family i, either MZ or DZ), and ϵ^θ an individual specific n-dimensional environmental zero-mean shock on the skill. We specifically denote the effect of family income, which is assumed to be linear with coefficient π.

The skill of twin ij is thus given by [12] :

$$\theta_j^i = w(g_j^i) + \pi y^i + \Pi X_j^i + F^i + \epsilon_j^{\theta,i}. \tag{4}$$

We assume the no-correlation:

$$\forall i, j, \forall k \in \{h, y\} : \mathbf{E}\epsilon_j^{k,i}\epsilon_j^{\theta,i} = 0, \tag{5}$$

and zero-mean conditions:

$$\forall i, j : \mathbf{E}F^i = 0, \mathbf{E}(\epsilon_j^{\theta,i}) = 0.$$

2.4. **Parental Investment.** The i^{th} household solves in the variables E expenditure in consumption and I^i pair of investment in the two children:

$$\max_{(E^i, I_1^i, I_2^i)} \mathbf{E}_{(\theta_1^i, \theta_2^i)} \left((1-\delta) \ln E^i + \delta \sum_{j=1,2} y_j^i \right), \tag{6}$$

[11] We ignore the possible measurement error of PGS here, since we are not primarily interested in heritability per se. A possibe extension of our research would reduce this attenuation using for example methods described in Becker et al. (2021).

[12] The effect of family income on skill in equation (4) is taken here as given. One can easily set up a more complex model on which parents also decide an investment in skill formation, in addition to human capital accumulation as described in the next section 2.4. This more complex model, is described in section S-0.1 of the Appendix. where we show that it yields a skill equation just like (4), and where the income term is produced by a household optimization problem, just as it is in equation (8).

subject to the budget constraint given by the household's income (recall y is the natural log of income):

$$E^i + \sum_{j=1,2} I_j^i = \exp(y^i) \tag{7}$$

The choice on consumption and educational investment is taken with the knowledge of the skills (θ_1^i, θ_2^i) of the children, hence the sub-script in the expectation of equation (6), which refers to the random shocks ϵ^h and ϵ^y. Human capital accumulation is described by:

$$h_j^i = \alpha_I \ln I_j^i + \alpha_\theta \theta_j^i + \epsilon_j^{h,i}, j = 1, 2 \tag{8}$$

and income is given by:

$$y_j^i = \alpha_h h_j^i + \epsilon_j^{y,i}, j = 1, 2.^{13} \tag{9}$$

We assume zero mean for shocks to human capital and income:

$$\forall i, j, \forall k \in \{h, y\} : \mathbf{E}\epsilon_j^{k,i} = 0; \tag{10}$$

and assume that the shocks to human capital and income are not correlated:

$$\forall i, j : \mathbf{E}(\epsilon_j^{h,i} \epsilon_j^{y,i}) = 0;$$

At the optimal solution of the problem in equations (6-10), 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:

$$\hat{I}^i = \frac{\delta \alpha_{Ih}}{1 - \delta + 2\delta\alpha_{Ih}} \exp(y^i) \equiv \psi \exp(y^i). \tag{11}$$

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

2.5. **Income of the Children.** In the analysis below we also use this more general model to control for education of parents, college degree of parents, work status of the father. Substituting the optimal investment reported in equation (11) into the human capital equation (8) and substituting the result into equation for income (9) we get the reduced equation for income:

$$y_j^i = a + \alpha_{Ih} y^i + \alpha_{\theta h} \theta_j^i + \alpha_h \epsilon_j^{h,i} + \epsilon_j^{y,i} \tag{12}$$

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

To complete the model, we need to specify how the pairs of parents are selected. To this we turn now.

[13] In both equations (8) and (9), we could add on the right hand side a term $w(g)$, multiplied by some additional parameter, to allow direct influence of genetic component on the variable. However, since this term already appears in the right hand side of equation (4), this genetic component will, even in the simple version presented in equations (8) and (9), be considered in empirical estimates, and this addition would make the model more complex with no substantial gain.)

2.6. **Matching Processes.** To complete the system described by equations (4), (8), (12), (18) and (19), 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, namely 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 the personality traits, different from cognitive or non-cognitive skills, that are recorded in our data). Recall Y is the set of log-incomes, let $Z \equiv G \times Y \times \Theta \times C$, and the observable characteristics $Z_O \equiv Y \times \Theta \times C$ with generic element z_O; for convenience we indicate with a subscript (as in $\Delta_m(Z)$) whether the element in $\Delta(Z)$ refers to the mother 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 denoted $M(\mu_m, \mu_f) \equiv \nu \in \Delta(Z \times Z)$, 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 of each type of parent distribution is the same as the original distribution for that type:

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

(2) **Conditional independence of genotype:** the random variables g_m and g_f (genotype of mother and father) are independent, conditionally on the information of observable characteristics.

The conditional independence assumption requires that matching only depends on the observable characteristics $z_O \in Y \times \Theta \times C$; in other words, matches are made on the basis of observable characteristic and not on the genotype. Thus, matching of genotypes is not random within the population, but it is random within the set of individuals with given observable characteristics. The assumption is very weak, at least as long as individuals choose their partners without taking into account the results of genetic tests, which is typically not yet the case.

Random Matching within the entire population is a special example of matching: in this case, a mother of type z_{mO} is selected, and independently a z_{fO} for the father, according to μ_m and μ_f respectively. This model is convenient for its simplicity, but it is not entirely 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 will provide a better approximation. A detailed analysis of the equilibrium concept is presented in section A.1.

2.7. **Matching According to Worth.** The analysis of the invariant distribution is simpler if matching is only dependent on income and skill of the spouse. So we set:

$$\Pi = 0, F^i = 0, \epsilon^\theta = 0, \epsilon^h = 0, \epsilon^y \sim N(0, \sigma^2_{\epsilon^y}). \tag{13}$$

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).

A pair of genotype and income (g, y) has a worth $w(g) + w_y y$. Mating is random within each worth class. To define these classes, we consider partitions of the set. A possible partition is the *discrete partition*, in which mating occurs only within pairs of exactly the same worth; we will use this partition as a simple but not very realistic example. A more realistic model has a *countable partition*. To define it, we take a countable set of values, indexed by the integers:

$$\mathcal{V} \equiv \{v_i : i \in \mathbb{Z}\}. \tag{14}$$

We assume that these values are increasing in the index, and that the distance between successive terms is uniformly bounded above and below:

$$\exists \underline{M}, \overline{M}, \forall i : 0 < \underline{M} \leq v_{i+1} - v_i \leq \overline{M}. \tag{15}$$

The class of genotype and income pairs of worth v_i is defined as

$$C(v_i) \equiv \{(g, y) : w(g) + w_y y \in [v_i, v_{i+1})\}. \tag{16}$$

The worth function $W : G \times Y \to \mathcal{V}$ is defined as:

$$W(g, y) \equiv v_i \text{ if } (g, y) \in C(v_i). \tag{17}$$

We consider a probability measure $\mu \in \Delta(G \times Y, \mathcal{B}(G \times Y))$, where $\mathcal{B}$ are the Borel subsets, as the description of the current distribution in the population of pairs of genotype and income. G is finite, so the Borel σ-field is the power set; using the Borel definition and notation for both G and Y simplifies the exposition.

As we mentioned, children in our sample are all twins. The genetic transmission function in equation (1) is obviously true in particular for each individual twin. In addition to that equation we have two additional conditions restricting the joint transmission to the pair of twin. These conditions depend on the twin type, an element on the set $\{DZ, MZ\}$, and are defined as:

$$H_{DZ}(g_m, g_f)(g^1, g^2) = H(g_m, g_f)(g^1) H(g_m, g_f)(g^2) \tag{18}$$

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

$$\begin{aligned} H_{MZ}(g_m, g_f)(g^1, g^2) &= H(g_m, g_f)(g^1) \text{ if } g^1 = g^2 \\ &= 0 \text{ otherwise} \end{aligned} \tag{19}$$

for MZ twins.

For the given μ, we describe the next period measure as follows. Each worth class is chosen with the probability induced by μ on the worth space, denoted by μ_V. Two parents (that is, two pairs of genotype and income) are chosen according to the probability on that class of genotypes and income. Within the class, mating is random. The genotype of parents then determines genotype of the child, and parents' income and education, together

with child's genotype, determine income. This entire process yields the new measure. [14] The complete model of the process on genotype, income, education and skill is given by equations (4) for skill, (8) for education, the reduced equation (12) for income, and (18) and (19) for the genotype transmission, Together with the mating process presented in section 2.6, these equations completely determine a non-linear (because of the function H in equation (1)) transition on measures on the space of genotypes and income, $\Delta(G\times Y)$. An invariant distribution is a fixed point of this transition function.

If an invariant distribution exists, we can then subtract from the variables $(y^i_j, \theta^i_j, h^i_j, w(g^i_j))$ 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 (18) and (19) in the compact form:

$$g^i_j \text{ is distributed as } H_k(g^i_m, g^i_f),\ k \in \{MZ, DZ\}. \tag{20}$$

If we substitute equation (4) into the reduced equation for income (12) we get the twin's income y^i_j as a linear function of genetic endowment g^i_j, family income y^i and environment F^i, and a weighted sum of idiosyncratic (j dependent) variables:

$$\begin{aligned} y^i_j &= \alpha_{\theta h} w(g^i_j) + (\alpha_{Ih} + \alpha_{\theta h}\pi) y^i + \alpha_{\theta h} F^i + \\ &\quad \alpha_{\theta h} \Pi X^i_j + \alpha_{\theta h} \epsilon^{\theta,i}_j + \alpha_h \epsilon^{h,i}_j + \epsilon^{y,i}_j. \end{aligned} \tag{21}$$

The decomposition in equation (21) is a more detailed version of the standard ACE decomposition in behavioral genetics (see for example Knopik et al. (2017), page 358), where the phenotype is income, the A term is the additive contribution of genotype, $\alpha_{\theta h} w(g^i_j)$, the common or shared environment component C is the sum of the two terms $(\alpha_{Ih} + \alpha_{\theta h}\pi) y^i$ and $\alpha_{\theta h} F^i$, and sum of the last four terms is the E component.

We assume:

$$\alpha_{Ih} + \alpha_{\theta h}\pi < 1, \alpha_{\theta h} > 0, \tag{22}$$

to ensure that (the first inequality) the income process is bounded, and an invariant measure exists, and (the second inequality) that skill has a non trivial effect on income. The equation describing human capital accumulation is similar, up to the constant multiplier α_h; we report it here for convenience because we will cite it in the empirical analysis:

$$h^i_j = \alpha_\theta w(g^i_j) + (\alpha_I + \alpha_\theta \pi) y^i + \alpha_\theta \Pi X^i_j + \alpha_\theta F^i + \alpha_\theta \epsilon^{\theta,i}_j + \epsilon^{h,i}_j. \tag{23}$$

and is obtained substituting (11) into (8) and subtracting the constant term.

[14] For a precise definition of the transition from one period's measure to the next, we refer to section A.4; here the income of the child is described by equation (55), and genotype of the child by equation (57).

Different further specifications of the model are possible, depending on how we model the variables X_j^i and F^i in equation (23) and therefore in equation (21). We explore these possibilities in detail in the rest of the paper. In particular the equation modeling the variable F^i is examined in the section on passive gene-environment correlation (section 3.1); and the model for the variable X_j^i is analyzed in the section on measurement error 4.1, where we discuss how we plan to estimate equations (21) and (23), thus providing a link between theory and empirical analysis.

3. Invariant Measures

We now show that an invariant measure exists, and has some interesting properties. Existence of the invariant measure is far from immediate because the process on distributions of skills and income in our model is non-linear. The non-linearity follows from the matching process: in every period the two distributions (for potential mothers and fathers respectively) are shuffled by the matching to produce a measure on the product space of spousal pairs.

A few preliminaries are necessary for a good understanding of the statement. We call the *skill allele* at some locus the allele which yields a higher value of the skill (more precisely, it has a higher genic value). [15] We will find that, at equilibrium, matching is random within each worth class, thus alleles are in Hardy-Weinberg equilibrium at all loci, but the frequency may differ across classes. We recall that a population is in *Hardy-Weinberg equilibrium* at a bi-allelic locus (with alleles denoted A and a, and frequency of A equal to p) if the frequency of the three combinations (aa, aA, AA) are respectively $((1-p)^2, 2p(1-p), p^2)$; these are the combinations obtained by independent combination of two gametes carrying A or a (one from the father and one from the mother) with probability p and $1-p$ respectively. Under some assumptions (described in detail, for example, in section 3.1 of Nagylaki (1992) or section 2.2 of Crow and Kimura (1970)), and in particular the assumption that mating among male and female is random, a Hardy-Weinberg equilibrium is reached in one generation, and maintained in all following generations. Finally, recall that we assume (equation (22)) that skill affects income, but the total coefficient of household's income on children's is less than 1. We can state:

Theorem 3.1. *Assume (22), and that the worth of an individual depends linearly on income and skill. Then for any vector of allele frequencies:*

1. *An invariant measure exists, which induces that allele frequency;*
2. *Within each worth class, alleles at each locus are in Hardy-Weinberg equilibrium;*
3. *Within each worth class of the discrete partition, a higher income of both parents implies a lower expected polygenic score of the child;*

[15] 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)).

(4) *The allele frequencies are invariant across periods.*

Some remarks may help to clarify the statement. An invariant measure exists in spite of the process being non-linear, and for any initial allele frequency. The proof relies on the order structure of the genotype and income space. The Hardy-Weinberg equilibrium holds, but only within worth classes. One can thus compute the fixation index, which is a measure of populations differentiation due to genetic structure across populations (in our model, populations are income and skill classes). The deviations from Hardy-Weinberg in the population may be small, since the phenotype is highly polygenic, and the size of $GWAS$ coefficient declines quickly. Still, as we are going to see in section 3.2.2, the model predicts a stratification across populations of the alleles with stronger effect. Higher income of both parents is compensated by the lower skill implicit in the genotype (the third statement). The last statement shows that frequency in the population of each allele does not change from one period to the next. So there may be many invariant measures depending on the initial condition (at least 2^K, see theorem A.4). The intuitive reason for the invariance property is that, as long as income does not affect the relative fertility for different genotype and income, the specific features of the mating process may affect the association of genotype and income, but can only reshuffle the existing alleles. The lack of differential effect on fertility is a strong assumption, particularly when we are interested in secular development, and examining the implications of relaxing it is an essential next step in research.

3.1. **Gene-Environment Correlation.** In our estimation (presented in sections 5) of the two equations (21) and (23) we will consider among the independent variables the polygenic score of the parents. We justify here the reason for this choice. Clearly, all the information on the genotype of the parents that could be potentially 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. However, 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*, rGE; Plomin et al. (1977), Scarr and McCartney (1983), Jaffee and Price (2007)).

The idea of *Gene-Environment correlation* (usually denoted rGE) rejects the assumption that environment and genes are uncorrelated. [16] The correlation may arise in three main ways. The most important for our purposes is the *passive* rGE effect. [17] 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 G and E. For example, higher intelligence of parents, due in part to the genes of the parents, may be transferred directly through genes to children, but also through the family environment created by parents. A related concept, *genetic nurture* has been extensively explored in Kong et al. (2018) and Okbay et al. (2022), and we discuss it below. [18]

We now discuss how rGE can be analyzed within our model, and how we can then estimate it in our data analysis. First, the household income (y^i in equation (23)) is already an example of an rGE path: the income of the parents is determined in part by their genes (this follows applying the income equation (21) to the parents) and also by the grand-parents' genes (iterating the process) and so on. Similarly, if we include among the variables in the vector X_j^i the human capital h^i of the parents, then applying the human capital equation (23) to the parents, and iterating, we see that parents', grandparents' genes and so on are relevant. Since the entire ancestry of the individual enters into the determination of the family income and parents' education, we refer to this as *ancestral* rGE. Models of parental investment as in Becker and Tomes (1979) are a special, very simplified, case of ancestral rGE. We have information of family income and parents' education in our data, and so we can control for its effects. But passive rGE may arise in a different more subtle way, which we model by considering the case in which,

[16] rGE is different from *Gene-Environment interaction* (usually denoted $G \times E$). The latter 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. For example, in a 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.

[17] 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 in which to operate by different genetic types. These two effects are harder to estimate in our data.

[18] Genetic nurture in Kong et al. (2018) is defined to operate through those genes that are not transmitted from parents to children. The role of family environment in considered in detail in Willoughby et al. (2021), which is discussed in detail in section 7.

in equation (23), the variabke F has the special form:

$$F^i = \alpha_m^C g_m^i + \alpha_f^C g_f^i \tag{24}$$

that is, the family environment depends on the genetic profile of the parents though some k dimensional vectors that may differ for father and mother. The additive form is the same we make for the genes affecting directly educational attainment. In section A.5 of the appendix we provide a detailed analysis of this case.

We emphasize that the weights α^C in equation (24) may be very different from those estimated by β in section 2.3. In particular very different genes (more precisely, SNP's) can be relevant in equations (3) and (24). We provide an example of this difference below, where the two sets of genes are disjoint. We also emphasize that "parents" in equation (24) should be interpreted in the more precise meaning of individuals providing parents' role. For example, if the child is adopted then the genotypes (g_m^i, g_f^i) in equation (24) are those of the adopting parents, not the biological ones (and the same holds for y^i and h^i). With minor changes, the proof of theorem 3.1 holds, and thus in particular existence of an invariant measure holds. We will refer to this component of rGE as *parental* rGE.

3.2. **Numerical computation.** The main properties of the process and equilibrium distribution of the model in section 3 can be illustrated with a numerical computation of the equilibrium distribution. [19] We study the distribution in $\Delta(G \times Y)$ in successive generations of a constant size population where each household has two children. The sex of each child is determined independently (from each other and from the other variables) with probability 1/2 on each sex.

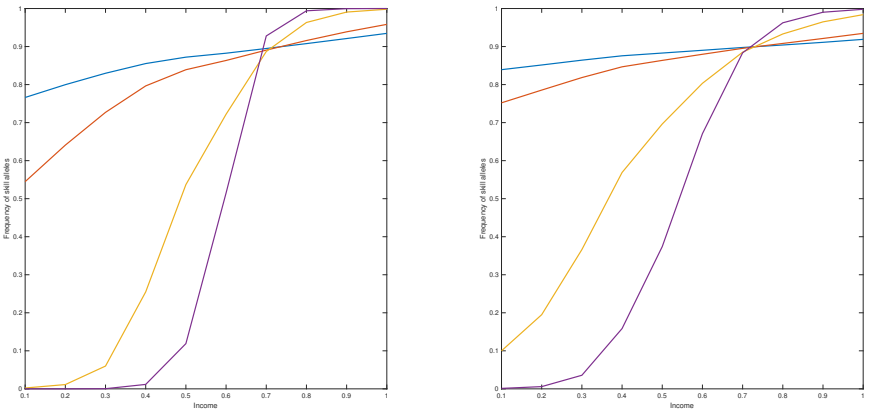
3.2.1. *Speed of Convergence.* Convergence to the invariant distribution is fast, and approximately achieved in our model within five generations. The value of the ratio of the norm of the difference between current and past μ, and the norm of the current μ is within ten per cent after five generations, and 2.26 per cent after ten generations.

3.2.2. *Endogenous Population Stratification.* The skill alleles have at equilibrium a frequency that is increasing with worth, education and income. As we mentioned in theorem 3.1, society is stratified. The effect is strong, and is stronger the higher the genic value of the allele. Both facts are illustrated in the left panel of figure 1.

3.2.3. *Parental* rGE. An intuitive reason for the next result is provided by a simplified example. Consider the case in which the set of genes (or more precisely the SNP's) that are relevant in equation (3), and the other set of those relevant for equation (24), have empty intersection. We refer to the first set as EA (for educational attainment) SNP's and to the second as

[19] Coding in Matlab (R2022b). The Matlab code is available upon request.

FIGURE 1. **Population Stratification and** rGE. Both panels display the frequency of alleles by income. The flattest line, with smallest difference across income, describes the frequency of the allele with smallest genic value; the others are in increasing order, with purple line for the highest effect allele. *Left Panel:* Only child's genotype affects income (no passive rGE). *Right Panel:* Only parents' genotypes affects income (full rGE). The figure illustrates how two very different economies may have very similar statistical properties.



PC (for parental care). SNP's improving parental care also affect positively education and income of the children. Obviously, children's parental care SNP's are correlated with those of the biological parents, by 50 per cent or more (due to assortative mating); and since parents' PC SNP's affect educational attainment of the children, these SNP's will be correlated to educational attainment, and thus will appear to influence educational attainment directly even if they are not. This is illustrated by the comparison of the top and bottom panels of figure 1. The two sets of panels report main features of two economies that have the same underlying preferences and technology, but completely different pathways from genes to traits; that is they only differ in the functions w and F. We will refer to the economy in the left panel as the EA economy and that in the right panel PC economy for short. [20]

[20] The bottom panels of the figure rely on the analysis developed in section A.5. In the notation of that section, there is a K dimensional vector α such that, in the top panels $\alpha^A = \alpha, \alpha_m^C = \alpha_m^C = 0$; and in the bottom panels, $\alpha^A = 0, \alpha_m^C = \alpha_m^C = \frac{1}{2}\alpha$. In simple words, the left panel describe an economy where all alleles are EA, and no passive rGE exists; in the right panel, no allele affects educational attainment, and the effect is only through the environment provided by the parents.

The figure illustrates the following results. First, just as in the case in which the effect on educational attainment is direct, also when the genetic effect occurs only through parental care there is population stratification, with higher frequency of the alleles with positive effect in the richer, more educated population (see bottom left panel). Second, for each allele k the distribution of income for the three subgroups of the population with $g(k) = 0$, 1 or 2 is different, even in the economy where there is no direct genetic effect on education (see bottom right panel). Third, as a consequence of the second point, the estimated $GWAS$ coefficient for educational attainment are significant and positive even in this latter economy, where there is no direct effect of genes on educational attainment.

Obviously, in principle parental rGE affects children's phenotype. The real question is: once we control for ancestral rGE, is parental rGE quantitatively important, once we control for the ancestral one? In section (7) we show that the answer is negative.

3.3. **Inter-generational mobility: standard and genetic model.** In this section we compare the predictions of the model we have presented with those of the standard model of parental investment. The model with autoregressive transmission of skill (as introduced in Becker and Tomes (1979)) has (adopting our notation to this case) the following equations for income in generation t:

$$y_{t+1} = \alpha_{Ih} y_t + \alpha_{\theta h} \theta_{t+1} + \epsilon^y_{t+1} \tag{25}$$

and for skill:

$$\theta_{t+1} = \eta \theta_t + \epsilon^\theta_{t+1} \tag{26}$$

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 ρ_{PM} (the subscript PM stands for Perfect Matching; the reason for this will be clear in the comments following equation (32) below. will be soon clear) to be:

$$\rho_{PM} = \alpha_{Ih} + \alpha_{\theta h} \frac{\eta \mathbf{E}(\theta y)}{\mathbf{V}(y)} \tag{27}$$

where $\mathbf{V}$ denotes the variance of a random variable, and $\mathbf{E}(\theta y)$ and $\mathbf{V}(y)$ have an explicit expression in terms of the primitive parameters. [21] When

[21] The explicit expressions are:

$$\mathbf{V}(\theta) = \frac{\sigma^2_{\epsilon^\theta}}{1 - \eta^2} \tag{28}$$

$$\mathbf{E}(\theta y) = \frac{\alpha_{\theta h} \mathbf{V}(\theta)}{1 - \alpha_{Ih} \eta} \tag{29}$$

$$\mathbf{V}(y) = \frac{1}{1 - \alpha^2_{Ih}} \left(\alpha^2_{\theta h} \mathbf{V}(\theta) + \sigma^2_{\epsilon^y} + 2 \alpha_{Ih} \alpha_{\theta h} \eta \mathbf{E}(\theta y) \right) \tag{30}$$

$\sigma_{\epsilon^y} = 0$, the inter-generational persistence formula (27) becomes the well known formula (see e.g. Solon (2004)) in which persistence is a simple weighted average of the income and skill transmission:

$$\rho_{PM} = \frac{\alpha_{Ih} + \eta}{1 + \alpha_{Ih}\eta} \tag{31}$$

A direct comparison of the standard model (equations (25) and (26)) with a genetic model like (20) and (21), 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 A.1 and 2.7. We assume income and skill to be the weighted average of the income and skill of the two parents, as in equations (49) and (50). Thus, the income of the child follows the equation:

$$y_{t+1} = \alpha_{Ih} \sum_{i=m,f} w_i^y y_{it} + \alpha_{\theta h}\theta_{t+1} + \epsilon_{t+1}^y \tag{32}$$

and the skill transmission follows:

$$\theta_{t+1} = \eta \sum_{i=m,f} w_i^\theta \theta_{it} + \epsilon_{t+1}^\theta \tag{33}$$

The matching between parents that decides the pairing of (θ_{mt}, y_{mt}) with (θ_{ft}, y_{ft}) is determined by preferences and stable matching as in section A.1. The standard model (25 - 26) becomes a special case of (32 - 33) when we assume that preferences of mothers and fathers are lexicographic (with any order on θ 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 (27) or (31) of the standard model are an *upper bound* on the persistence within the class of models requiring equations (25), (32) and (33). 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 ρ_{MY}, can be shown to satisfy:

$$\rho_{MY} < \rho_{PM} \tag{34}$$

The proof is in section A. 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 η (in equation 26). Such a fixed parameter, however, has no correspondent in reality: the genetic model shows that the persistence represented by that η is endogenous, and depends on

the distribution of the genotype. Therefore the corresponding elasticity, call it ρ_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 (presented in equations 32) and (33)). 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 (32 –33)) 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 (52); and to ease comparison with the simple form (31) we assume $\sigma_{\epsilon^y} = 0, \Pi = 0, F = 0, \epsilon^\theta = 0$.

This economy has a stationary distribution at two values:

$$(0, y_0, \theta_0) \text{ with prob } 1 - p(A), (2, y_2, \theta_2) \text{ with prob } p(A), \tag{35}$$

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 increasing the variance of genetic values, and magnifying the heritability and the resemblance between relatives. [22] This effect is absent by assumption in the autoregressive model, even in the amended version on which two partners are introduced, given by equations (25), (32) and (33).

4. Estimation Strategy

Our empirical analysis will estimate equations (21) and (23). In the next two subsections we discuss the introduction into the analysis of the genotype

[22] 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 multivariable example. The analysis in population genetics is very different form the one we present here because the assortative mating in our model is endogenous and determined at equilibrium in the marriage market.

of the parents among the explanatory variables (X_j^i) and the additional information we can derive from a special subset of our data, the DZ twins. We recall that the joint distribution of genotypes is described by equations (18) and (19)

4.0.1. *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 types of 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 Θ exist. [23]

Whatever the cause, the correlation for DZ twins is a simple function of the correlation between the PGS of the parents. We use the subscripts $1, 2$ to indicate that the variable refers to first and second sibling; and subscripts m and f, for mother and father respectively. Then:

Lemma 4.1. *The correlation between the standardized PGS of non identical full siblings, hence in particular of DZ twins, is equal to* $\frac{1}{2}$ *plus half of the correlation between the standardized PGS of the parents, that is:*

$$\mathbf{E}(PGS_1 PGS_2) = \frac{1}{2} + \frac{1}{2}\mathbf{E}(PGS_m PGS_f))$$

The proof is in section A.3. [24] Lemma 4.1 gives the predicted correlation among DZ twins as a function of the correlation among parents. In section 7 below we present the correlation among parents' PGS, and find that data are consistent with the prediction of the lemma.

In the next sections we will test and estimate the parameters of the two equations (21) for income and (23) for human capital. The data we use are described in detail in section 9.

[23] We can illustrate this second possibility considering the extreme case in which there is no overlap between loci affecting the θ 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 θ, thus different distributions on the θ skills. This equilibrium is not robust, of course: with a small imperfection in the C-matching the frequency of the θ 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.

[24] On the related, but different, issue of segregation variance (that is, the variance of the offspring about the mid-parent value), see Rogers (1983).

4.1. **Measurement Error and SEM.** Reliable estimates, for example that of the path from genetic factors to educational outcomes, must take into account errors in measurement of the variables. This is obviously important if we want to minimize downward biases of single coefficients; but it is even more important if we want to compare the *relative* size of the effect operating through cognitive and non cognitive skills, since the error in measurement might be different for the two groups of variables. For example, it might be natural to expect a larger error in measures of non cognitive skills, based on surveys, than in cognitive, based on tests. We model explicitly and estimate errors in measurement using a structural equation model, (SEM).

The SEM we consider is of the usual (see for example Bollen (1989)) form:

$$\mathbf{Y} = B\mathbf{Y} + \Gamma\mathbf{X} + \alpha_Y + \zeta \tag{36}$$

with $\mathbf{Y}$ a m-vector of m_Y endogenous observed variables, $\mathbf{y}$, and m_η endogenous unobserved variables η; $\mathbf{X}$ an n-vector of exogenous n_x observed variables $\mathbf{x}$ and n_ξ exogenous unobserved variables ξ, α_Y a vector of means, and ζ a vector of errors. Entries of B are denoted by β's, entries of Γ by γ's. In this section we adopt the notation convention that variables with capital first letter are endogenous, and lower case first letter are endogenous. [25]

We set up our analysis adopting a general form (36) to test the basic equations of the model, with basic equations (21) and (23). Specific examples are the system of equations (39) – (40) and that of equations (43) – (45). We recall that the variables in the vector X_j^i for ij (introduced in section 2.3) are not necessarily observed, so we add equations providing measurements of these latent variables. They may also be endogenous, so we add equations describing how they are determined. Examples of variables which are components of X_j^i are the latent endogenous variables C and NC (cognitive and non-cognitive skills, in equations (37) and (38)), (these are the η-variables); observed endogenous variables e_h and y_h in equations (43) and (44), ($\mathbf{y}$-variables); and finally pGS_m and pGS_f, exogenous observable $\mathbf{x}$-variables in equations (43) – (45).

5. Income and Human Capital Determination

5.1. **Income Analysis.** We first estimate the parameters of the model presented in section 2.7. Table 1 below reports the panel regression of the log income at the age 29 take over family income, PGS, and other control variables. Estimates reported in the table control for the difference in the age of the individuals (parent or child) at which the information on income was collected. Since wage increases with age at a rate that may be heterogeneous (Rupert and Zanella (2015), Lagakos et al. (2018)), this difference may introduce a bias in the estimated coefficient, if the slope depends on

[25] This carries the modest price of changing PGS to pGS

characteristics like education that are correlated with wage. We use a specification of the Mincer (1974) equation which has that of Lagakos et al. (2018), [26] as special case, allowing the slope to depend on education. If the slope increases with education, we expect the estimated elasticity coefficient to overestimate the true value; thus we control for the time difference, education of the parents, and an interaction term.

TABLE 1. **Income at the age 29 take, family income, PGS, and Personality.** All variables, including College of parents and Male, are standardized to mean zero and SD 1. The signs of MPQ variables NA, Externalizing and Academic problems are reversed. Controlled for PC's and the parents-child time difference in age at income data collection.

	(1) b/se	(2) b/se	(3) b/se
Family Income	0.134***	0.128***	0.078**
	(0.027)	(0.027)	(0.032)
Male	0.277***	0.276***	0.313***
	(0.025)	(0.025)	(0.029)
Male × Family Income	–0.060**	–0.060**	–0.050*
	(0.025)	(0.025)	(0.030)
PGS		0.078***	0.021
		(0.025)	(0.028)
Education Years			0.256***
			(0.035)
IQ			0.008
			(0.029)
MPQ PA			0.061**
			(0.026)
MPQ NA			–0.024
			(0.027)
MPQ CN			0.034
			(0.032)
Externalizing			–0.072*
			(0.037)
Academic effort			0.057
			(0.038)
Academic problems			–0.017
			(0.034)
N	2100	2100	1485

The estimated unconditional inter-generational elasticity (IGE) is 0.134 (SE = 0.027); the table (Model (1)) reports the values after control for sex and interaction between sex and family income. Age has not a significant effect, as might be expected since since individuals in the sample are approximately

[26] Specifically the formulation given in section VI A.

the same age. Sex of the individual has a strong and significant effect: income for male individuals has a substantially larger intercept (27.7 per cent), but a smaller (by 6 per cent) dependence on the family income. The fraction of males in the twins population is 48 per cent; thus, the standardized male variable is approximately equal to 1 for male and -1 for female.

In Model (2), the coefficient of the individual polygenic score is 7.8 per cent (SE $= 0.025$, p-value $= 0.002$). Its size is approximately half of that of family income (12.8 per cent). Considering that the polygenic score we are using is estimated from coefficients from a GWAS for education, it is likely that the weight of genetic factors affecting income is higher.

Model (3) in the table presents controls for some of the variables that are likely to mediate the effect of the polygenic score. Education Years is the most natural variable to capture the effect of the polygenic score in education, and in fact the estimated coefficicent is large (25.6 per cent, (SE $= 0.035$), p-value < 0.001) and significant. [27]

The controls for principal components and the difference in the age of parents and children at the collection of data on income produce no significant coefficient; the IGE falls after the control for difference in age, as expected, but in small measure (in the order of 10 per cent). Controls for additional variables (in particular Education of Parents, Polygenic score of father and mother) produces elasticity coefficients that are small and non significant, with no effect on the coefficients of the variables of more significant interest. [28]

The values of IGE are on the lower side of the currently available estimates for developed countries, which vary between a minimum of 0.2 and 0.4 (see for example Zimmerman (1992), Solon (1992), Lee and Solon (2009), Mazumder (2005); and surveys in Björklund et al. (2012) and Blanden (2011)). The coefficient reaches higher values in some studies: see for example Palomino et al. (2018) who in a finer analysis (taking into account quartiles of the distribution), show it can take larger values for the highest and lowest levels of income. There are some possible explanations for this difference. One is measurement error in our income data. Another is that in some developed countries with European population the IGE coefficient is lower. For example, in Sweden (a country that is more relevant given the demographic composition of Minnesota at the time in which the data were collected) values are lower (see for example Österberg (2000) where values are around 0.125 (page 427)); although it can be substantially higher at higher values of income (see Björklund et al. (2012)) which are less relevant for our sample.[29]

[27] Note that the sample size is smaller because several variables are missing for some subjects.

[28] The coefficients are: 0.004, (SE $= 0.166$) for parents' education, -0.03, (SE $= 0.037$) for mother's PGS, and 0.04, (SE $= 0.038$) for father's PGS.

[29] See Björklund and Jäntti (1997) for a comparison of Sweden and USA on intergenerational mobility who find mobility higher in Sweden.

6. Identifying the path from PGS to education

In this section we identify how much of the effect of PGS on educational achievement can be attributed to factors such as cognitive or non cognitive skills.

In our estimates below the vector $\mathbf{Y}$ has a vector η of endogenous latent variables equal to (C, NC), denoting cognitive and non-cognitive skills respectively. The structural observed component of the $\mathbf{Y}$ vector is number of education years of the twins, e: we focus of this measure (rather than college, or GPA) because it is the most relevant for economic consequences. The measurement variables in $\mathbf{Y}$ are a vector of $((ct_i)_{i=1,...,I_{ct}}), (nct_j)_{j=1,...,J_{nct}})$ of measurements of cognitive and non-cognitive skills. Turning to the vector $\mathbf{X}$, in our case $x \equiv ((x_k)_{k=1,...,K})$ is a vector of control variables, such as household income variables, education and polygenic score of parents, the principal components, age and sex. The system we estimate is:

$$C = \beta_C \, pGS + \zeta_C \tag{37}$$

$$NC = \beta_{NC} \, pGS + \zeta_{NC} \tag{38}$$

$$ct_i = \alpha_{ct_i} + \gamma^C_{ct_i} C + \zeta_{ct_i}, i = 1, \ldots, I_{ct} \tag{39}$$

$$nct_j = \alpha_{nct_j} + \gamma^{NC}_{nct_j} NC + \zeta_{nct_i}, j = 1, \ldots, J_{nct} \tag{40}$$

$$e = \alpha_e + \gamma^C_e C + \gamma^{NC}_e NC + \sum_k \gamma^{x_k}_e x_k + \zeta_e. \tag{41}$$

$$\gamma^C_{ct_1} = \gamma^{NC}_{nct_1} = 1. \tag{42}$$

The PGS may be added to the right hand side of the equation (41) with little consequence. The normalization condition (42) is necessary because any multiplication of the variables β_C and ζ_C by a positive constant, and corresponding division by the same constant of the vector $(\gamma_{ct_i} : i = 1, \ldots, I_{ct})$ gives a new vector of parameters, with the corresponding random variables still satisfying the system of identification equation; a similar re-scaling of β_N, ζ_N and $(\gamma_{nct_i} : i = 1, \ldots, J_{ct})$ would have the same effect. Hence the two normalization conditions (42). With this normalization, the model is identified, if there are at least two cognitive and two non-cognitive tests. More precisely:

Proposition 6.1. *Assume $I_{ct} \geq 2$ and $J_{nct} \geq 2$, then the system (37)-(42) is identified.*

Proof. Substituting equations (37) and (38) into the equations (39) - (41) reduces the system to a system of observed variables. We indicate by σ^2_X the variance of a variable X. The simpler system in observed variables can be solved recursively for the parameters in the following order: σ^2_{pGS}, β_C, β_{NC}, $\gamma^C_{ct_i}$, $\gamma^{NC}_{nct_i}$, $\sigma^2_{\zeta_C}$, $\sigma^2_{\zeta_{NC}}$, $\sigma^2_{\zeta_{ct_i}}$ for all $i \neq 1$, $\sigma^2_{\zeta_{nct_j}}$ for all $j \neq 1$, γ^C_e, γ^{NC}_e and finally $\sigma^2_{\zeta_e}$. □

TABLE 2. *SEM* **of Pathways from PGS to Education Years.** The model estimated is described in equations (37) to (41). All observed variables standardized to mean zero and SD 1. Cognitive skills test scores (*ct*'s) are verbal and performance IQ, non-cognitive (*nct*'s) are the three broad *MPQ* dimensions. Standard errors estimated by bootstrapping. $N = 852$. Model *vs* saturated: $Pr > \chi^2 < 0.0001$.

Equation	Variable	b	z	p value	CI
Ed Yrs	C	0.285	4.87	<0.001	[0.171, 0.401]
		(0.058)			
	NC	0.856	3.11	0.002	[0.315 , 1.4397]
		(0.276)			
	PGS	0.014	0.35	0.725	[-0.066 , 0.94]
		(0.041)			
	PGS mother	0.033	0.71	0.282	[-0.027 , 0.093]
		(0.030)			
	PGS father	0.019	0.66	0.512	[-0.039 , 0.078]
		(0.030)			
	Educ Parents	0.136	4.58	<0.001	[0.078 , 0.194]
		(0.29)			
	Family Income	0.075	2.38	0.017	[0.013 , 0.137]
		(0.031)			
	Male	-0.151	-2.77	0.007	[-0.260 , -0.041]
		(0.055)			
	Constant	0.376	9.85	<0.001	[0.301 , 0.450]
		(0.027)			
C	PGS	0.287	9.21	<0.001	[0.226,0.349]
		(0.031)			
NC	PGS	0.040	1.95	0.051	[-0.0002, 0.081]
		(0.025)			

The structural component of the *SEM* estimation is reported in table 2. In the equation for education years, the coefficient both C and NC are significant. We can compute with the delta method the product of the coefficient for the link from the PGS to the variable C, times the coefficient from C to Education Years. The value of the product is 0.082, (SE = 0.018, $z = 4.53$, p-value < 0.001), with confidence interval $[0.046, 0.117]$. The corresponding product for the path passing through NC has a value of 0.034, (SE = 0.019, $z = 1.8$, p-value= 0.071), with confidence interval $[-0.003, 0.072]$.

Once we control for C and NC, the coefficient of the PGS is not significant (p-value $= 0.725$). For comparison we note that in the regression restricted to twins, controlling only for sex, the coefficient is 18.7 per cent (SE $= 0.022$, $z = 8.37$, p-value < 0.001). The coefficients of education of parents and family income are both significant, of the same order of magnitude, but education of parents (13.6 per cent (SE $= 0.029$, $z = 4.58$, p-value < 0.001)) is approximately twice that of family income (7.5 per cent (SE $= 0.031$, $z = 2.38$, p-value $= 0.017$). The PGS of parents is not significant.

7. Fixed Effects Analysis and parental rGE

In this section we estimate the equations for income and human capital using two important additional pieces of information: the fact that children are twins, both DZ and MZ, the overlapping generation structure of the model, and the information, including the genetic one, on parents. We begin with the analysis based on DZ twins.

7.1. **Fixed Effects Analysis with 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.

We have seen in section 4.0.1 the theoretical estimate of the correlation among DZ twins depending on the degree of assortative mating of the parents. The difference in PGS correlation and the predicted correlation with random assortative matching (which is $\frac{1}{2}$) is 0.083 and it must be due to the assortative matching among parents. In our case we are considering not the genome-wide correlation [30] but the one between PGS of parents. The correlation coefficient between PGS of the two parents is $r = 0.152$). As discussed recently in the 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 (PC's) is performed. [31] In Table S-10 of the appendix (section S-0.5) we report the controls for PC's in our data. The table 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: for education years, the correlation coefficient is $r = 0.522$, for IQ is 0.37.

[30] See Robinson et al. (2017), Supplementary Note, page 12.

[31] See Section S2 Principal Components of Domingue et al. (2014), Table S1. These are the same tests we use in Table S-10.

The fixed effects analysis is presented in tables S-2, S-3, S-4, S-5, for education years, GPA, college and IQ score respectively. All the regressions show that the coefficient of the PGS is significant in the fixed effects regression. In the case of GPA the coefficient is large and is approximately equal in the two regressions.

7.2. **The Explanatory Power of Parents' PGS.** A different way to control for the effect of genetic endowment of parents on family environment is to control directly for their PGS; in this case we can use the information on both types of twins, including MZ.

The system we estimate is presented in equations (43) to (45), and is a special case of the general SEM structure in the general system (36), with the same interpretation for the parameters α_Y, β, γ and σ as in section 4.1. As usual, the superscript i refers to the family, and the subscript j to the twin. The variables e_h and y_h denote education of parents (average of the education years of the two parents) and family income. The $\mathbf{y}$ variables are (e_h, y_h, e); the $\mathbf{x}$ variables are (pGS_m, pGS_f, pGS). There are no exogenous latent ξ-variables. The equations of the model are: Note that differently from the estimate reported in table 2, we are not controlling for C and NC variables. Our model is:

$$e_h^i = \alpha_{e_h} + \gamma_{e_h}^{pGS_m} pGS_m^i + \gamma_{e_h}^{pGS_f} pGS_f^i + \zeta_{e_h}; \tag{43}$$

$$y_h^i = \alpha_{y_h} + \gamma_{y_h}^{pGS_m} pGS_m^i + \gamma_{y_h}^{pGS_f} pGS_f^i + \zeta_{y_h}; \tag{44}$$

$$e_j^i = \alpha_e + \gamma_e^{e_h} e_h^i + \gamma_e^{y_h} y_h^i + \gamma_e^{pGS} pGS_j^i + \gamma_e^{pGS_m} pGS_m^i + \gamma_e^{pGS_f} pGS_f^i + \zeta_e. \tag{45}$$

With this formulation we can take into account the difference in polygenic score of the DZ twins, and still use the information on MZ twins (see Okbay et al. (2022) for a justification of this method). The estimate of the SEM model is presented in table 3.

Education of parents and family income have a strong and significant influence on educational attainment of the twins, thus they exert their influence though this channel in addition to the direct one of the genotype of the twins. However, the coefficients of the two parental polygenic scores, which could potential report additional unobserved channels from genotype of parents to education years, are not significant, although they are of course large and significant in the equations for both family income and parents' education. This finding is consistent with the result reported in Willoughby et al. (2021): conditioning on parental IQ and socioeconomic status substantially reduces the effect of parental genotype. Within our model, this

result is an implication of the identification of family income and parental education [32] as the pathways of the effect of family background. [33]

The results are similar if we introduce explicitly a latent variable F of family environment, affected by the PG of the parents, and modify the education years equation as:

$$e_j^i = \alpha_e + \gamma_e^{e_h} e_h^i + \gamma_e^{y_h} y_h^i + \gamma_e^F F^i + \gamma_e^{pGS} pGS_j^i + \zeta_e. \tag{46}$$

PGS of parents significantly affect education of parents and income of the family; and in turn education of parents and income of the family affect education years of children, but F has little residual influence.

[32]These variables were not considered not considered Willoughby et al. (2021).

[33]For the record, the coefficient of the score of the mother is significant at the 10 per cent level.

TABLE 3. *SEM* **of Pathways from PGS to Education Years.** The model estimated is described in equations (43) to (45). All observed variables standardized to mean zero and SD 1. Standard errors estimated by bootstrapping. $N = 802$. Model *vs* saturated $Pr > \chi^2 < 0.0001$:

Equation	Variable	b/se	z	p value	CI
Educ Parents	PGS mother	0.182	5.62	<0.001	[0.118 0.245]
		(0.032)			
	PGS father	0.301	8.96	<0.001	[0.235, 0.367]
		(0.033)			
	Constant	0.066	2.00	0.045	[0.001, 0.132]
		(0.033)			
Family Income	PGS mother	0.091	3.12	<0.001	[0.034, 0.149]
		(0.029)			
	PGS father	0.154	5.05	<0.001	[0.094, 0.213]
		(0.030)			
	Constant	0.131	4.28	<0.001	[0.070, 0.198]
		(0.030)			
Ed Years	Educ Parents	0.183	8.76	<0.001	[0.142, 0.224]
		(0.021)			
	Family Income	0.112	4.84	<0.001	[0.066, 0.157]
		(0.023)			
	PGS	0.103	4.84	0.002	[0.038, 0.167]
		(0.032)			
	PGS mother	0.052	2.26	0.094	[-0.006, 0.084]
		(0.023)			
	PGS father	-0.003	-0.13	0.899	[-0.051, 0.044]
		(0.024)			
	Male	-0.139	-2.85	0.004	[-0.235, -0.043]
		(0.048)			
	Constant	0.345	13.43	<0.001	[0.284, 0.395]
		(0.025)			

We find similar results if we consider different measures of educational attainment. For example, if we take the variable e_j^i to be a binary variable indicating whether the twin has a college degree or not (and estimate the correspondent of (45) with a probit model) we find the coefficients of e_h to be 0.35 (SE $= 0.05$, $z = 6.98$, p value < 0.001; marginal effect 12 per cent); for y_h the coefficient is 0.21 (SE $= 0.057$, $z = 3.76$, p value < 0.001; marginal effect 6.7 per cent). The estimated coefficient for the polygenic score of the twin is 0.16 (SE $= 0.043$, $z = 3.67$, p value < 0.001; marginal effect 5.4 per cent).

7.3. **Regression on Parents' PGS.** In this section we see that if we regress variables of interest on the polygenic score of the children and we include that of the parents, we typically find the coefficient of the parents' score to be significant and positive. This finding provides evidence that the genes of parents affect success of children in addition to the direct effect on the genes of the children. After we control for education of parents and family income, the coefficient of the parents' PGS is insignificant, while the coefficient of the PGS of the twin stays significant. This second finding suggests that income and education of parents channel most of the additional effect of parents' genes.

We present the results in section S-0.4 for education years (table S-6), GPA (S-7), college (S-8) and intelligence (S-9). These results are consistent with earlier findings of passive rGE, [34] but add insight in the mechanism from genetic profile of parents to children's outcomes: most of this effect is channeled by parents' education and income, with parents' education typically the largest and most significant.

When we control for education of parents and family income (see model (4) in equation S-6), the coefficients of the PGS of the parents is substantially reduced and not significant. In this model the fraction explained by education of parents is large (coefficient is 0.116, ($SE = 0.025$), and so is the case for family income (coefficient is 0.083, ($SE = 0.028$)). Interestingly, the coefficient of the mother's polygenic score shows some modest effect in models (2) and (3), that is even after we condition for IQ and soft skills. The same result of the decline of decline of significance of the PGS of parents holds for other indicators of educational attainment, such as college and the GPA index, reported in tables S-7 and S-8.

In conclusion, we add two findings to the analysis in Kong et al. (2018) and Willoughby et al. (2021), where evidence of a passive gene-environment correlation is reported. Fist, we identify, consistently with the model we developed in the theory section 2, and with the more general theory of parental investment, two paths through which genetic factors of the parents operate indirectly, namely family income and education of parents; education of parents with a larger coefficient than family income. Second, we show that

[34] See Kong et al. (2018), see also the analysis in Willoughby et al. (2021).

once these two factors are taken into account, there is no significant residual indirect effect. [35]

8. Conclusions

Our analysis has been setup as a natural extension of theories of parental investment and intergenerational mobility (as in Becker and Tomes (1979) and in the large 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 non-random mating equilibrium. Our model provides the basis for an economic analysis of genetic factors in education and intergenerational mobility; it is more realistic than the existing models, and it is still analytically manageable so that it can be tested in the data. Our data analysis provides a proof of concept of this statement.

Realism of the assumptions would matter little, perhaps, if the predictions of the alternative models were similar. We have shown instead that 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 ultimately on 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) though the environment provided by parents to children. If we want to analyze precisely the relative weight of nature and nurture, an issue which is crucial for a variety of public policies, economic theory will need to adopt models that incorporate this information explicitly. The difference between standard and fully specified genetic models will become even more consequential as more precise estimate of the link between genes and phenotypes of economic interest, as well as richer information on the genetic profile of individuals, become available.

In our empirical analysis we confirm earlier results that 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 of the score to about 60 per cent in the top decile. Very different pathways of the effect of PGS could be consistent with this finding: for example, the effect might be entirely due do discrimination operating on individual characteristics that are genetically based but irrelevant for the

[35] Within the model defined precisely here, there is little evidence of genetic nurture, as defined in recent literature (see for an in depth discussion Wang et al. (2021), Okbay et al. (2022)).

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 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 the effect of genotype of parents on the environment of children (passive rGE). Our analysis decomposes this effect into two different paths: one operates through genes that affect directly educational attainment of the parents, but influence the environment of the children indirectly through the effects on income ad education. This first is the path that economists have analyzed with model of parental investment. A second path operate trough genes that affect directly the environment of the children without affecting educational attainment of the parents, and thus their income and education. Our analysis of the data suggest that most of passive rGE operate though the first channel; within this channel, education matters more than income.

Fixed effects analysis on DZ twins is performed exploiting the information we have on the genotype, summarized by the polygenic score, which is identical for MZ twins and differs among DZ twins, in a measure that depends on chance and the degree of assortative mating between the child's parents. Our results 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. The analysis of the pathways operating from genes associated with educational attainment though cognitive and non cognitive skills show that the largest effect is through cognitive skills. [36]

9. Description of the Data

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 % (see McGue et al. (2014)).

[36] This conclusion is different from the one reached in McGue et al. (2020) and McGue et al. (2017), using the same data. The reason for the discrepancy is *ex-post* clear. Both these papers do not set up the analysis as a test of a fully specified model of parental investment, and ignore key variables in the analysis, such as household income.

9.1. **Measures of Income.** Data on income of parents and twins were collected at different points in time. The age of parents at the moment in which the data on income was collected is higher than the age of the children by approximately 10 years. We control for this difference in the estimation (see the discussion preceding table 1). The measure of Parents' Income was collected on a 13-point, self-report scale that ranged from 1 = less than $10,000 to 13 = Over $80,000. [37]

A first assessment of 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?". No specific band of income was suggested. In the analysis the data on income are translated into dollar amount, then log transformed, and standardized.

9.2. **Measures of Human Capital.** Information on educational achievement in the sample is provided by a classification of the individual in seven classes, described in Table 4. 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 .8.

TABLE 4. **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

[37] The precise 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.

9.3. **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.

Computation of PGS. We constructed the polygenic scores predicting years of education from the summary statistics released by Lee et al. (2018), with the cohorts 23andMe and MTFS removed. The weights of the SNP's in the score were then calculated with the software tool LDpred (Vilhjálmsson et al. (2015)), which uses an external sample to estimate the correlations between SNP's in order to convert the univariate regressions coefficients in $GWAS$ summary statistics to partial regression coefficients. We used the data in MCTFR parents of European ancestry to estimate the correlations between SNP's and calculated the partial regression coefficients of the 450,000 SNP's that were originally genotyped in MCTFR and survived all default software filters. We set the LDpred shrinkage parameter equal to unity–the highest possible value and the one leading to the least shrinkage of the PGS weights. This choice, sometimes regarded as the most conservative, was followed by Lee et al. (2018). Our experience has shown that varying this parameter over a tenfold range scarcely influences the prediction R^2 (e.g., Willoughby et al. (2021)).

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 of Affectivity (here PA, associated with Wellbeing, Social Potency, Achievement, and Social Closeness), Negative Emotionality or Affectivity (NA, 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. [38]

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

[38]For details, see `https://www.upress.umn.edu/test-division/mpq/copy_of_mpq_BF-overview`.

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 (with $\alpha = .91$) [39] 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").

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.

[39] Cronbach's alpha (Cronbach (1951)) is a good lower bound on the reliability when the scale measures only one common factor

Appendix A. Proofs and Additional Material

A.1. **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 $\Theta \equiv \times_{l=1}^{n} \Theta_k$, each component has a natural order (such as "taller", "more intelligent", "lower Neuroticism score" and so on), and Y has the natural order over the real numbers, so Θ and $\Theta \times Y$ have an 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:

$$\forall z_M^{''}, z_M^{'} : z_M^{''} \geq z_M^{'} \text{ implies } \forall c \in C, (z_M^{''}, c) \succeq_{z_{Os}} (z_M^{'}, c) \tag{47}$$

and homophilic:

$$\forall z_M, c, e, f : d(f, c) \leq d(e, c) \text{ implies } \forall z_M^{'}(z_M^{'}, f) \succeq_{(z_M, c)_s} (z_M^{'}, e). \tag{48}$$

The household maximization problem described in equation (6)-(10), which only depends on the $\Theta \times Y$ components, defines a preference over matches. In the maximization problem an individual (θ_m, y_m) evaluates the utility $U(\theta_m, y_m, \theta_f, y_f)$ from a match with an individual (θ_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 θ^c is linear combination of the skills of the parents with weight w_i^{θ}, $i \in \{m, f\}$ also adding to 1. In summary we assume:

$$y^h = w_m^y y_m + w_f^y y_f; \tag{49}$$

and

$$\theta^c = w_m^{\theta} \theta_m + w_f^{\theta} \theta_f; \tag{50}$$

Substituting the optimal investment (11) into the budget constraint (7)), the education (8) and income(9) equations we find that, up to a constant independent of θ and y, the *worth* in the marriage market of a type (θ, y) of sex $i \in \{m, f\}$ is:

$$W_i(\theta, y) \equiv (1 - \delta + 2\delta\alpha_{hI}) w_i^y y + 2\delta\alpha_{h\theta} w_i^{\theta} \theta \tag{51}$$

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

$$U(\theta_m, y_m, \theta_f, y_f) = W_m(\theta_m, y_m) + W_f(\theta_f, y_f) \tag{52}$$

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.[40] By the properties we have derived we conclude using standard arguments:

Proposition A.1. *A stable matching exists. There is complete segregation over C. Parents' genotypes (the random variables g_m and g_f) are conditionally independent for any vector of observable characteristics.*

A.2. **Proof of inequality 34.** The inequality follows because when parents match on income and only on income the system (28-30) is as follows. Equation 28 becomes:

$$\mathbf{V}(\theta) = \frac{\sigma^2_{\epsilon^\theta} + \frac{\eta^2}{2}\mathbf{E}(\theta_m, \theta_f)}{1 - \frac{\eta^2}{2}}. \tag{53}$$

Equations (29) and (30) are unchanged. Rearranging one obtains the inequality 34.

A.3. **Proof of lemma 4.1.** We denote PGS_i the polygenic score of twin i, and PGS_m, PGS_f, as indicated by the subscript the score of mother and father respectively. SImilar notation for g, the genotype of various individuals. The proof uses the fact the the genotype of the child is (after meiotic recombination) the sum of one haplotype of the mother and one of the father, each chosen with equal probability. Recall we are considering an additive model, as stated in equation (2). Given these premises, we have:

$$E(PGS_i|g_m, g_f) = \frac{PGS_m + PGS_f}{2} \tag{54}$$

We then have:

$$\begin{aligned}
& \mathbf{E}(PGS_1 PGS_2) \\
= \; & \mathbf{E}(\mathbf{E}(PGS_1 PGS_2)|g_m, g_f) \\
= \; & \mathbf{E}(\mathbf{E}(PGS_1)|g_m, g_f)\mathbf{E}(PGS_2)|g_m, g_f)) \\
= \; & \mathbf{E}(\mathbf{E}(\tfrac{1}{2}(PGS_m + PGS_f))|g_m, g_f)\mathbf{E}(\tfrac{1}{2}(PGS_m + PGS_f))|g_m, g_f)) \\
= \; & \tfrac{1}{2}\mathbf{E}(\mathbf{E}((PGS_m)^2 + PGS_m PGS_f))|g_m, g_f) \\
= \; & \tfrac{1}{2} + \tfrac{1}{2}\mathbf{E}(PGS_m PGS_f))
\end{aligned}$$

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 (equation (54)), the fourth from symmetry between PGS_m and PGS_f, and fifth again follows from elementary properties of expectation.

[40] More precisely: A matching ν is stable, if and only if for all, except possibly a zero measure set (with respect to the product measure $\nu \otimes \nu$), pairs $(z_m, z_f, z_m^{'}, z_f^{'})$,

$$z_f \succ_{z_m} z_f^{'} \text{ or } z_m^{'} \succ_{z_f^{'}} z_m \text{ or } \left(z_f \succeq_{z_m} z_f^{'} \;\&\; z_m^{'} \succeq_{z_f^{'}} z_m \right).$$

A.4. **Proof of Theorem 3.1.** We recall the equations describing the process on income and genetic profile, simplifying the notation for clarity in exposition.

We write the income equation in the compact form:

$$y_c = \beta C(y_m, y_f) + w(g_c) + \sigma Z \tag{55}$$

where $\beta < 1$, $\sigma > 0$, Z is a standard normal, and C is a composition map giving a household income as function of the income of the two parents. We assume that C is continuous and satisfies:

$$\min\{y_m, y_f\} \leq C(y_m, y_f) \leq \max\{y_m, y_f\}; C(y, y) = y. \tag{56}$$

We will call *household income*, and denote it y_h, the value of this composition. This form includes the special cases in which the household income is the average of the two parents' income, possibly with different weights.

The genotype of the child given the pair of parents' genotypes (g_m, g_f) is a random variable with distribution, conditional on (g_m, g_f)

$$H(\cdot|g_m, g_f) \in \Delta(G). \tag{57}$$

We can now be more precise. We first assign to μ its disintegration according to the partition $W^{-1}(\mathcal{V})$, that is the vector of pairs of probability of the class v_i and the conditional probability given v_i:

$$((\mu_{\mathcal{V}}(v_i), \mu(\cdot|v_i)) : i \in \mathbb{Z}), \tag{58}$$

By Rohlin's theorem (Rohlin (1952)), such a disintegration exists and in addition (i) $\mu_{\mathcal{V}}$ is a probability measure on $\mathcal{V}$, (ii) for every i, $\mu(\cdot|v_i))$ a probability measure on $G \times Y$ that satisfies $\mu(C(v_i|v_i)) = 1$ and (iii) $\mu(\cdot) = \sum_{i\in\mathbb{Z}} \mu_{\mathcal{V}}(v_i)\mu(\cdot|v_i))$.

We now describe the function giving the next period measure, examining each component of this object separately. First there is a Markov kernel assigning to a parents' profile (g_m, y_m, g_f, y_f) a probability on $G \times Y$, interpreted as the child's genotype and household income, assigning to $O_G \times O_Y \in \mathcal{B}(G) \times \mathcal{B}(Y)$

$$K_F((g_m, y_m, g_f, y_f), O_G \times O_Y) = H(O_G|g_m, g_f)\delta_{C(y_m,y_f)}(O_Y) \tag{59}$$

Next is the Markov kernel assigning to a pair (g_c, y_h) of child's genotype and household income a distribution on child's income, as by equation (55), assigning to a Borel subset of Y, O_Y:

$$K_I((g_c, y_h), O_Y) \equiv \Phi\left(\frac{1}{\sigma}(O_Y - \beta y_h - w(g_c))\right) \tag{60}$$

where Φ is the measure induced by the standard normal. We define the function $\Psi : \Delta(G \times Y) \to \Delta(G \times Y)$ as:

$$\Psi(\rho) \equiv (\rho \otimes \rho)K_F K_I \tag{61}$$

where $\rho \otimes \rho$ is the independent product of the two measures, and $K_F K_I$ is the composition of the two kernels.

Lemma A.2. *The map Ψ is continuous in the weak topology.*

Proof. The map $\rho \to \rho\otimes\rho$ is continuous (see Lemma 1.1, chapter 3 of Parthasarathy (1967)). The rest follows from the continuity assumption on the combination function C and the fact that the topology on G is discrete, thus H is continuous. □

The next period measure is defined by the function

$$T : \Delta(G \times Y, \mathcal{B}(G \times Y)) \to \Delta(G \times Y, \mathcal{B}(G \times Y)),$$

where for every set $O \in \mathcal{B}(G \times Y))$:

$$(T\mu)(O) \equiv \sum_{i \in \mathbb{N}} \mu_{\mathcal{V}}(v_i)\Psi(\mu(\cdot|v_i))(O) \tag{62}$$

As standard in economics, we will study the distribution on population characteristics (genotype and income) considering the invariant distributions.

A.4.1. *Invariant Measures.* This section will illustrate a reason why the model with a fully specified genetic transmission is different from the standard model.

The following invariance property is true for any function T' (including the function T we defined earlier) on $\Delta(G \times Y)$ that has two basic properties. The first is the *mating property*: the mating process operates though a mating function $M : (\Delta(G \times Y))^2 \to \Delta((G \times Y)^2)$ that preserves marginals. In the case of T defined in equation (62), the mating function is:

$$dM(\mu, \mu) = \sum_{v_i} \mu_V(v_i) \left(\mu(\cdot|v_i) \otimes \mu(\cdot|v_i)\right)$$

The second is the *factor property*: the distribution of child's genotype and income factor through the H function in equation (57) and a kernel $S : (G \times Y)^2 \to \Delta(Y)$ denoted $S(\cdot; (g_m, y_m, g_f, y_f))$ (in the case of T, this is the Markov kernel $K_F K_I$).

Lemma A.3. *The set of measures with the same minor allele frequency is invariant under any T' that satisfies the mating and the factor property.*

Proof. Let for this proof $C : G \to \{0, 0.5, 1\}^K$ defined by $C(g, k) \equiv g(k)/2$, and the push forward mapping $\mu \in \Delta(G)$ to $C_*\mu$; the expectation with respect to $C_*\mu$ at k gives the frequency of the allele at locus k as

$$AF(k) = \int_G d\mu_G(g)C(g, k)$$

Then, denoting $X \equiv (G \times Y)^2$, with generic element $x \equiv (g_m, y_m, g_f, y_f)$, the next period allele frequency at k is:

$$\begin{aligned}
&\int_{G\times Y} \int_X dM(\mu, \mu)(x)H(g_c; g_m, g_f)S(dy_c; x)C(g_c, k) = \\
&\int_G \int_X dM(\mu, \mu)(x)H(g_c; g_m, g_f) \int_Y S(dy_c; x)C(g_c, k) = \\
&\int_G \int_X dM(\mu, \mu)(x)H(g_c; g_m, g_f)C(g_c, k) = \\
&\int_G \int_{G^2} dM(\mu, \mu)_{G^2}(g_m, g_f)H(g_c; g_m, g_f)C(g_c, k) = \\
&\int_G d\mu_G(g)C(g, k)
\end{aligned}$$

where the first equality follows from Fubini's theorem; for the second we have used the obvious fact that for all x:

$$\int_Y S(dy_c; x) = 1;$$

for the third we have defined for $O \in \mathcal{B}(G^2)$

$$M(\mu,\mu)_{G^2}(O) = M(\mu,\mu)(O \times Y^2);$$

and the last follows from the basic properties of the function H. □

The following proposition examines a case which is uninteresting from a substantial point of view (because it excludes heterogeneity), but is very useful for illustration of the differences between our model and the standard model of parental skill transmission. Let us define the set of genotypes that are homozygotes at all loci:

$$Hom \equiv \{g \in G : \forall k, g(k) \in \{0,2\}\}, \tag{63}$$

a set of 2^K elements. If the marginal of the initial measure is concentrated on a single element in Hom, then all the iterates have the same property.

Proposition A.4. *The map T has at least 2^K fixed points.*

Proof. Take the initial measure to be concentrated on a single genotype $g \in Hom$. We consider for illustration the case in which the partition is fine. In the general case the result follows as a corollary of our results below. With the fine partition mating takes place among individuals with the same income and genotype. There is no dynamics involving G, so the is a unique invariant measure is distributed as $N\left(\frac{w(g)}{1-\beta}, \frac{\sigma^2}{1-\beta^2}\right)$. □

Note that the dynamic is entirely in the set $\Delta(Y)$; restricted to this set, the iterates of T are weakly asymptotically stable. Of course the initial condition is not, in the interesting case, nconcentrated on an element in Hom.

A.4.2. *Estimates of T.* As we mentioned in the text, the specific difficulty in analysing T derives from the fact that, due to the product of measures in the definition of Ψ, T is not linear. Thus standard theorems on existence of invariant measures, such as Krylov-Bogoliuvov which is based on averaging, are not available.

To address this difficulty we first endow $G \times Y$ with a partial order. We say that $g' \succeq_G g$ if $w(g') \geq w(g)$ and we define the partial order on $G \times Y$, denoted $\succeq$, as the one induced by the $\succeq_G$ and the natural order over the real numbers. The order $\succeq$ allows us to define the set of increasing functions on $G \times Y$ as:

$$\mathcal{I} \equiv \{f : G \times Y \to \mathbb{R}, (g',y') \succeq (g,y) \Rightarrow f(g',y') \geq f(g,y)\}. \tag{64}$$

In turn we can now define the first order stochastic dominance order on probability measures on $G \times Y$ as the stochastic order induced by the cone $\mathcal{I}$.

We can now construct our estimates of the function T. In simple terms, the idea is to construct a function that is defined by the same process on income and genotype as T is, but gives the best possible income and the best possible genotype to the child. This will give us a control from above, and a similar procedure will give the control from below. Since the construction for the lower bound is completely symmetric to that of the upper bound we will develop in detail only the first.

Our control from above will operate on the subset of measures that have support on the best possible genotype, that we now define. We let g^* and g_* to be any choice of g providing the maximum and minimum value, respectively, of the function w

on the finite set G, arbitrarily selecting one of the optimal values if necessary; that is:

$$\forall g \in G : w(g^*) \geq w(g) \geq w(g_*)$$

We will refer to g^* (g_*) as the selected best (worst) genotype. The first step is to define the largest class to which an income can belong, for some genotype:

$$V(y) \equiv \max\{v_i : G \times \{y\} \cap C(v_i) \neq \emptyset\} \tag{65}$$

Conditions (15) and (16) insure that the function V is well defined, that is, that the supremum is finite and it is achieved. Also note that by definitions (16) and (17), $V(y) = W(g^*, y)$; definition (65) is more convenient for future use. Next we define the sup over the incomes in a class:

$$\overline{Y}(v_i) \equiv \sup\{y : G \times \{y\} \cap C(v_i) \neq \emptyset\}. \tag{66}$$

Note that $W(g_*, \overline{Y}(v_i)) = v_{i+1}$.

Lemma A.5. *The function V is piecewise constant, increasing, and right-continuous. The function $y \to \overline{Y}(V(y))$:*

1. *is piecewise constant, increasing, and right-continuous;*
2. *is such that, for all $y \in Y$ such that $V(y) = v_i$,*

$$\frac{w(g^*) - w(g_*)}{w_y} \leq \overline{Y}(V(y)) - y \leq \frac{v_{i+1} - v_i + w(g^*) - w(g_*)}{w_y} \equiv y_{\overline{Q}}.$$

Proof. Let $B^* : \mathcal{V} \to Y$ be defined by

$$B^*(v_i) = \frac{v_i - w(g^*)}{w_y}.$$

Note that

$$(\{g^*\} \times Y) \cap C(v_i) = \{g^*\} \times [B^*(v_i), B^*(v_{i+1}))$$

The function V is constant and equal to v_i on the interval $[B^*(v_i), B^*(v_{i+1}))$, hence the statement concerning V follows. The function $\overline{Y}(V(\cdot))$ inherits the properties of V, so is piecewise constant and right continuous. The function $y \to \overline{Y}(V(y))$ has on the interval $[B^*(v_i), B^*(v_{i+1}))$ the minimum at $B^*(v_{i+1})$ and the maximum at $B^*(v_i)$, and the values in the statement follow from simple computations, with the difference $B^*(v_{i+1}) - B^*(v_i)$ providing the additional term $\frac{v_{i+1}-v_i}{w_y}$ in the upper estimate. □

We denote the subset of measures with full support on the selected best genotype:

$$\Delta^*(G \times Y) \equiv \{\nu \in \Delta(G \times Y) : \nu(\{g^*\} \times Y) = 1\} \tag{67}$$

Lemma A.6. *For $\mu \in \Delta(G \times Y)$ and $\nu \in \Delta^*(G \times Y)$,*

$$\nu \succeq \mu \text{ if and only if } \nu_Y \succeq \mu_Y$$

Proof. If $\nu \succeq \mu$, then considering functions that are constant with respect to G proves that $\nu_Y \succeq \mu_Y$. If $\nu_Y \succeq \mu_Y$, then for any $h \in \mathcal{I}$,

$$\begin{aligned}(\nu, h) &= \int_Y d\nu(g^*, y)h(g^*, y) \\ &\geq \int_Y d\mu(g, y)h(g^*, y) \\ &\geq \int_Y d\mu(g, y)h(g, y)) \\ &= (\mu, h)\end{aligned}$$

where the first equality is the definition, the second is the hypothesis we made, the third follows because $h \in \mathcal{I}$, and the last is the definition. □

We now introduce the function on measures that will provide the upper bound for T; it is denoted $\overline{Q}$ and we provide first a description of its definition. Take any y, and assign to both parents the income $y + y_{\overline{Q}}$, so $y_h = y + y_{\overline{Q}}$, and genotype g^*. Then apply the same transition from pair of parents' genotype and income as we do for T. The induced function on measures $\overline{Q}$ is linear.

Definition A.7. *The Markov kernel $S^{\overline{Q}}$ is defined as, for any $O_Y \in \mathcal{B}(Y)$:*

$$S^{\overline{Q}}(y, O_Y) \equiv \Phi\left(\frac{1}{\sigma}(O_Y - \beta(y + y_{\overline{Q}}) - w(g^*))\right)$$

The function $\overline{Q}$ from $\Delta^*(G \times Y)$ to itself is defined as

$$(\overline{Q}\nu)(\{g^*\} \times O_Y) = \int_Y d\nu(\{g^*\}, y)S^{\overline{Q}}(y, O_Y) \tag{68}$$

Lemma A.5 implies that any household income obtained by a match in the class $V(y)$ is less than $y + y_{\overline{Q}}$; since any genotype g_c obtained by that match has $w(g_c) \leq w(g^*)$ the next period income obtained by this process dominates in first order stochastic dominance that induced by the process underlying T. Thus, for every y and μ in the order interval:

$$S^{\overline{Q}}(y, \cdot) \succeq S^T_\mu(y, \cdot)$$

where $\succeq$ is the order on operators (see chapter 5, second part of definition 5.2.1 in Müller and Stoyan (2002); see also O'Brien (1975), Kamae et al. (1977)).

We also recall that the sequence of iterates $P^n, n \in \mathbb{N}$ of a Markov operator on a metric space X is called weakly asymptotically stable if P has a unique invariant distribution μ^* and

$$\forall \mu \in \Delta(X, \mathcal{B}) : P^n \text{ converges weakly to } \mu^*.$$

Lemma A.8. *The function $\overline{Q}$ has a unique fixed point, $\overline{\nu}^\infty$, given by:*

$$\overline{\nu}^\infty(\{g^*\}, \cdot) \sim N\left(\frac{\beta y_{\overline{Q}} + w(g^*)}{1 - \beta.}, \frac{\sigma^2}{1 - \beta^2}\right). \tag{69}$$

The sequence of its iterates is weakly asymptotically stable.

Proof. Take the moment generating function of the n^{th} iterate of function defining the next period income random variable:

$$y' = \beta(y + y_{\overline{Q}}) + w(g^*) + \sigma Z$$

and consider the limit. □

To allow the comparison between T and $\overline{Q}$ we represent the action of T in a form similar to equation (68) for $\overline{Q}$. Since T is not linear, the Markov kernel corresponding to $S^{\overline{Q}}$ must depend on the current measure, and will be written as $S^T_\mu(y, O_Y)$ as the probability of a Borel set O_Y at the point y and population measure μ.

We first provide an informal description of the process underlying this special Markov kernel. The income of the parent m is chosen (this will be chosen according to the measure μ_Y). The genotype g_m is then chosen according to a version of the conditional measure $\mu(\cdot|y_m)$. The parent belongs to the class of worth $v_i = W(g_m, y_m)$, and a mate is chosen randomly in that class, with probability $\mu(\cdot|v_i)$. The parents' profile (g_m, y_m, g_f, y_f) gives the probability on child's pair (g_c, y_c).

The precise definition is given next:

Definition A.9. *For $\mu \in \Delta(G \times Y)$, $S^T_\mu : Y \to \Delta(Y, \mathcal{B}(Y))$ is defined as*

$$S^T_\mu(y, O_Y) \equiv \int_{G^2\times(G\times Y)} d\mu(g_m|y) \sum_{v_i} \delta_{v_i}(W(g_m, y)) d\mu(g_f, y_f|v_i) \tag{70}$$

$$H(g^c|g_m, g_f)\Phi\left(\frac{1}{\sigma}(O_Y - \beta C(y_m, y_f) - w(g_c))\right)$$

for any $O_Y \in \mathcal{B}(Y)$.

The Y-marginal of $T\mu$ is an average of $S^T_\mu(y, \cdot)$:

Lemma A.10. *For all $\mu \in \Delta(G \times Y)$ and $O_Y \in \mathcal{B}(Y)$:*

$$(T\mu)(G \times O_Y) = \int_Y d\mu_Y(y) S^T_\mu(y, O_Y) \tag{71}$$

Proof. We first observe that for any $\mu \in \Delta(G \times Y)$,

$$\sum_{v_i} \mu_V(v_i)(\mu(\cdot|v_i) \otimes \mu(\cdot|v_i))((g_m, y_m, g_f, y_f)) = \tag{72}$$
$$\mu_Y(y_m) d\mu(g_m|y_m) \sum_{v_i} d\mu(g_f, y_f|v_i)\delta_{v_i}(W(g_m, y))$$

Take now any real valued bounded continuous function f on Y:

$$
\begin{aligned}
&(T\mu, f) = \\
&\int_{G\times Y} d(T\mu)(g,y) f(y_c) = \\
&\int_Y f(y_c) \int_G d(T\mu)(g,y) = \\
&\int_Y f(y_c) \int_G \int_{(G\times Y)^2} \sum_{v_i} \mu_V(v_i)(\mu(\cdot|v_i)\otimes\mu(\cdot|v_i))((g_m,y_m,g_f,y_f)) \\
&H(g_c|g_m,g_f)Pr(y_c|y_m,y_f,g_c) = \\
&\int_Y f(y_c) \int_{(G\times Y)^2} \int_G \mu_Y(y_m) d\mu(g_m|y_m) \sum_{v_i} d\mu(g_f,y_f)|v_i)\delta_{v_i}(W(g_m,y)) \\
&H(g_c|g_m,g_f)Pr(y_c|y_m,y_f,g_c) = \\
&\int_Y d\mu_Y(y) \int_Y S^T_\mu(y,dy_c) f(y_c).
\end{aligned}
$$

where in the fourth equality we have used the initial observation (72); the second follows because f only depends on y, the third is the definition of T, and the last is the definition of $S^T_\mu(y,\cdot)$. □

We define the function $\underline{Q}$, the set $\Delta_*(G\times Y)$, the kernel $S^{\underline{Q}}$, measure $\underline{\nu}^\infty$, in a manner similar to $\overline{Q}$, $\Delta^*(G\times Y)$, $S^{\overline{Q}}$, $\overline{\nu}^\infty$ respectively.

We can now define the order interval

$$
[\underline{\nu}^\infty, \overline{\nu}^\infty] \equiv \{\mu : \underline{\nu}^\infty \preceq \mu \preceq \overline{\nu}^\infty\} \tag{73}
$$

Lemma A.11. *For every $\mu \in$, $T\mu \in [\underline{\nu}^\infty, \overline{\nu}^\infty]$.*

Proof. For μ in the order interval,

$$
\begin{aligned}
T\mu &\preceq \overline{Q}\mu \\
&\preceq \overline{Q}\overline{\nu}^\infty \\
&= \overline{\nu}^\infty
\end{aligned}
$$

where the first relation follows from $T \preceq \overline{Q}$, the second from monotonicity of $\overline{Q}$ (first part of definition 5.2.1 in Müller and Stoyan (2002)), and the last because $\overline{\nu}^\infty$ is a fixed point of $\overline{Q}$. □

The order interval has a key property, proved in the next lemma:

Lemma A.12. *The set $[\underline{\nu}^\infty, \overline{\nu}^\infty]$ is weakly compact, convex.*

Proof. Convexity is clear. We first prove that the set is relatively compact in the weak topology. By Prohorov's theorem (Parthasarathy (1967)), it suffices to show that it is uniformly tight. Let $\epsilon > 0$ be given: we claim that there exists a compact set $K \subseteq G\times Y$ such that for any μ in the set, $\mu(K) \geq 1-\epsilon$. We will find a set $K = G\times[-M,M]$ for some M. For such a K, $\mu(K) = \mu_Y([-M,M])$. By lemma A.6 we derive that

$$
\underline{\nu}^\infty_Y \preceq \mu_Y \preceq \overline{\nu}^\infty_Y.
$$

Find M large enough so that

$$\max\{\underline{\nu}_Y^\infty(-\infty,-M],\overline{\nu}_Y^\infty[M,+\infty)\}<\frac{\epsilon}{2},$$

so that

$$\mu_Y([-M,M]^c)<\epsilon$$

as required.

Finally, the order interval is weakly closed (see for example, Proposition 3 of Kamae et al. (1977)). □

Lemma A.13. *The function T on $[\underline{\nu}^\infty,\overline{\nu}^\infty]$ is continuous in the weak topology.*

Proof. The measures in the set are uniformly absolutely continuous with respect to the Lebesgue measure by the equation (55); note that the variance σ is independent of the income. Recall now that a sequence μ^n converges weakly to μ if and only if

$$\lim_{n\to\infty}\mu^n(A)=\mu(A)$$

for any Borel set A whose topological boundary ∂A has μ measure zero. Now the statement follows from the fact that for any $i\in\mathcal{Z}$

$$\partial C(v_i)=\cup_g\{(g,\frac{v_i-w(g)}{w_y}),(g,\frac{v_{i+1}-w(g)}{w_y})\}$$

which is a set of finite points in $G\times Y$. □

A simple example shows that continuity may fail when the uniform absolute continuity with respect to the Lebesgue measure fails.

Example A.14. *Let $K=1$, $G\equiv\{aa,aA,AA\}$, $w(aa)=0$, $w(aA)=1$, $w(AA)=2$, $w_y=1$. Let $v_1=0$, and $\mathcal{V}\equiv\{v_1\}$. Denote $(G\times Y)\setminus C(v_1)\equiv C(v_0)$, and denote the conditioning on the set $C(v_0)$ as conditioning on v_0.*

Consider the sequence in $\Delta(G\times Y)$:

$$\mu^n=\frac{1}{2}\left(p^n\delta_{(aa,\frac{1}{n})}+(1-p^n)\delta_{(AA,-2+\frac{1}{n})}\right)+\frac{1}{2}\left((1-p^n)\delta_{(aa,-\frac{1}{n})}+p^n\delta_{(AA,-2-\frac{1}{n})}\right) \tag{74}$$

with $p^n=\frac{2}{3}$ if n is even and $\frac{1}{3}$ when odd. If we also let:

$$\mu=\frac{1}{2}\delta_{(aa,0)}+\frac{1}{2}\delta_{(AA,-2)}$$

then μ^n converges weakly to μ.

We now consider the disintegration of the measures. For any n:

$$\mu_V^n(v_1)\equiv\mu^n(C(v_1))=\mu^n(C(v_0))=\frac{1}{2},$$

but

$$\mu(C(v_1))=1,\mu(C(v_0))=0$$

Also

$$\mu^n(\cdot|v_1)=p^n\delta_{(aa,\frac{1}{n})}+(1-p^n)\delta_{(AA,-2+\frac{1}{n})}.$$

and

$$\mu^n(\cdot|v_0)=(1-p^n)\delta_{(aa,-\frac{1}{n})}+p^n\delta_{(AA,-2-\frac{1}{n})}$$

On the other hand, $\mu(\cdot|v_0)$ is undefined and:

$$\mu(\cdot|v_1) = \frac{1}{2}\delta_{(aa,0)} + \frac{1}{2}\delta_{(AA,-2)}$$

Thus, the sequence of conditional expectations at a given worth oscillates with no limit, and the limit of any subsequence (when it exists) is different from the conditional value of the limit measure.

Also the function $\mu \to \mu_V(v_i)$ is not continuous.

We can now summarize the analysis developed so far, recalling the statement of theorem (3.1):

Theorem A.15. *Assume (22), and that the worth of an individual depends linearly on income and skill. Then for any vector of allele frequencies:*

1. *An invariant measure exists, with that allele frequency;*
2. *Within each worth class, alleles at each locus are in Hardy-Weinberg equilibrium;*
3. *Within each worth class of the discrete partition, a higher income of both parents implies a lower expected polygenic score of the child;*
4. *The allele frequencies are invariant across periods.*

Proof. This follows, given the previous analysis, from Himmelberg's theorem (Himmelberg (1972)).

The second part follows applying Hardy-Weinberg's theorem to the population within the worth class, and using the fact that equilibrium is reached in one period.

For the third part of the theorem, consider in the discrete partition case two families, indexed by $i = 1, 2$ with $y_j^2 > y_j^1$, $j \in \{m, f\}$, so the genotype worth, denoted w_j^i, is such that: $w_j^2 < w_j^1$, $j \in \{m, f\}$. The proof is very simple when the function w is injective. For any pair (w_m, w_f)

$$\begin{aligned}
&E(w(g_c)|w_m), w_f) \\
&= \sum_k \beta(k) E(g_c(k)|w_m, w_f) \\
&= \sum_k \beta(k) E(g_c(k)|g_m, g_f) \\
&= \sum_k \beta(k) E(g_c(k)|g_m(k), g_f(k)) \\
&= \sum_k \beta(k) \frac{1}{2}(g_m(k) + g_f(k)) \\
&= \frac{1}{2}\left(w(g_m)) + w(g_f)\right) \\
&= \frac{1}{2}\left(w_m + w_f\right).
\end{aligned}$$

Injectivity is used in the second equality. The third equality uses the absence of linkage disequilibrium among the SNP's in the polygenic score. In the general case in which $w^{-1}(w_j)$ is not a singleton, it suffices to take averages. Note the probability on the finite set $w^{-1}(w_m) \times w^{-1}(w_f)$ is uniform.

The last statement follows from lemma A.3. □

A.5. **Passive Gene × Environment Correlation.** We focus on triples of a child, mother and father, let $g_l^s \in \{0,1,2\}^K$ the genotype of $l \in \{c,m,f\}$, and with $s \in \{t, nt\}$, let $g_l^s \in \{0,1\}$ the transmitted ($s = t$) and non-transmitted part of the genotype of l. $g_l(k)$ and $g_l^s(k)$ are the values at the k^{th} locus. Note that

$$g_c = g_m^t + g_f^t, g_f = g_f^t + g_f^{nt}, g_m = g_m^t + g_m^{nt}. \tag{75}$$

We take α^A the 3^K-dimensional vector of true genic values of the genes as they affect directly the phenotype of interest (here A refers to the additive part in the standard ACE decomposition). α_l^C is the vector for the effect on the environment provided to the child by the parent of type l.

Recalling the form of the family environment variable in equation (24), and using equation (23), if we set $\Pi = 0$ to focus on the issue of interest, and take the value α_θ to be part of the genic values:

$$h_j^i = \alpha^A g_c + \rho y^i + \alpha_m^C g_m + \alpha_f^C g_f + \zeta_j^{h,i}. \tag{76}$$

where we have denoted, to lighten notation:

$$\rho \equiv \alpha_I + \alpha_\theta \pi, \alpha_\theta; \zeta_j^{h,i} \equiv \epsilon_j^{\theta,i} + \epsilon_j^{h,i}.$$

Equation (76) clarifies the different ways in which passive gene-environment interaction occurs. The first way is described by terms of the form $\alpha_l^C g_l$, which express the direct effect of the parents on the child's environment, through pathways that are possibly completely unrelated to the phenotype of interest (which is human capital in our case).

The second way operates through the term ρy^i, which contains implicitly terms of the form $\alpha_l^A g_l$, relative to parents, grandparents and so on, that affected the child's household income. Differently from the first, this pathway involves genes that are relevant for the phenotype of interest.

A.5.1. *Fully Genetic decomposition of income.* Recall that income is in our model a linear function of human capital with coefficient α_h. In the following we assume that we have rescaled the index of human capital so that $\alpha_h = 1$. We can now express the income of an individual as the discounted series of all past genetic contributions of ancestors, plus a random, zero mean term. To denote in a simple way the ancestors of an individual i we use the following notation. For any $n \in \{0,1,2,\dots\}$, a list of possible ancestors of depth n is an elements s in the set $\{m,f\}^n$. For instance, mi is the mother of i, fmi is the father of the mother of i and so on. We adopt the convention that at $n = 0$, the only element s in $\{m,f\}^n$ is the identity, so for such s, $si = i$, $smi = mi$ and so on. We denote with $h(i)$ the family of individual i.

Lemma A.16. *For every individual i:*

$$\begin{aligned} y_i &= \sum_{n=0}^{\infty} \left(\tfrac{\rho}{2}\right)^n \left(\sum_{s\in\{m,f\}^n} \left(\alpha^A g_{si} + \alpha_m^C g_{msi} + \alpha_f^C g_{fsi}\right)\right. \\ &+ \left.\sum_{s\in\{m,f\}^n} \zeta_{si}\right) \end{aligned} \tag{77}$$

Proof. Using (76) and recalling that $\alpha_h = 1$, we get for every individual i:

$$y_i = \rho y_{h(i)} + \alpha^A g_i + \alpha_m^C g_{mi} + \alpha_f^C g_{fi} + \zeta_i \tag{78}$$

where

$$y_{h(i)} = \frac{1}{2}(y_{mi} + y_{fi}). \tag{79}$$

Substituting equation (79) formulated for each ancestor repeatedly into (78) yields equation (77). The series converges under our assumption that $\rho < 1$. □

A.5.2. *Estimation.* Lemma A.16 has some useful implications for our estimations.

GWAS **coefficients.** The estimated *GWAS* coefficients $\beta(k)$'s of the k^{th} *SNP* are obtained as a linear univariate regression of the h^i_j values (or, given our normalization $\alpha_h = 1$, of y^i_j) on the $g_c(k)$ values. They are a biased estimate of the α^C values, for three reasons. The first reason is due to the term y^i in equation (76), because y^i is obviously correlated with g_c, since they are both affected by the parents' and other ancestors' genotype. The second factor is the term introduced by the environmental value F, given by the parents' genotypes, again correlated with g_c. The third factor is the Linkage Disequilibrium (LD) correlation between different loci.

We standardize the genotype variables to have mean zero and variance equal to (for the $g(k)$ variable) 1 and 1/2 (for the $g^s(k)$ variables), obtaining the new variables $Sg(k)$ and $Sg^s(k)$. [41] Using the formula in lemma A.16, if we ignore the LD correlation we find:

Lemma A.17. *For every k:*

$$\mathbf{E}\beta(k) = \alpha^A(k) + \frac{1}{2}\left(\alpha^C_m(k) + \alpha^C_f(k)\right) + \rho C \tag{82}$$

where C is a constant.

The term multiplied by ρ takes into account the effect occurring through grandparents and previous generations. As we have seen, ρ is between 0.2 and 0.4, thus terms with ρ or higher order are small. Eliminating the potential bias introduced by the terms of the form α^C is possible using information of the genotype of parents, direct or imputed (see Kong et al. (2018), Young et al. (2022)). Complete elimination of the bias would require information on the infinite sequence of ancestors, although the complete formula shows that the effects decays exponentially, thus effects of generations beyond parents is small.

We emphasize that, even if $\alpha^C_m = \alpha^C_f = 0$, a passive rGE effect persists through the influence on the environment of the children operating that genes influencing educational attainment produce on family income and parents' education. This effect may be substantial, and in our data it is. This what we consider next.

Controlling for parental *PGS*. We consider first the case with no effect of parents' genotype on environment, that is:

[41] That is, we call $p(k)$ the frequency of the allele with value 1, and define:

$$Sg(k) \equiv \frac{g(k) - 2p(k)}{\sqrt{2p(k)(1-p(k))}}; Sg^l(k) \equiv \frac{g^l(k) - p(k)}{\sqrt{2p(k)(1-p(k))}}, \text{ for } l = t, nt. \tag{80}$$

Of course at Hardy-Weinberg equilibrium:

$$\mathbf{E}Sg(k) = \mathbf{E}Sg^l(k) = 0, \mathbf{Var}Sg(k) = 1, \mathbf{Var}Sg^l(k) = 1/2, l = t, nt;$$

and

$$Sg(k) = Sg^t(k) + Sg^{nt}(k). \tag{81}$$

$$\alpha_m^C = \alpha_f^C = 0 \tag{83}$$

In this case, substituting equation (79) into (78) we obtain:

$$y_i = \alpha^A g_i + \frac{\rho}{2}\alpha^A(g_{mi} + g_{fi}) + \frac{\rho^2}{2}(y_{h(mi)} + y_{h(fi)}) + \zeta_i$$

The PGS of the child is an unbiased measure of the term $\alpha^A g_i$, and so are the parental scores for $\alpha^A g_{si}, s \in \{m, f\}$. Since ρ is relatively small, the larger part of the effect on income is produced by terms measured by the PGS of the child) and the PGS of the parents. This is the model we estimate in section S-0.4.

When assumption in equation 83 does not hold we have the bias described in equation (83), and at the current state of knowledge one has to accept it. However, the estimates presented in section S-0.4 suggests that adding the terms modeling the environmental effect changes little of the results.

Appendix B. Data availability

The data and codes necessary to replicate the empirical results in the paper are available at Harvard Dataverse, at the address

Rustichini, Aldo, William G. Iacono, James J. Lee and Matt McGue. "Replication Data for: 'Educational Attainment and Intergenerational Mobility: A Polygenic Score Analysis'," Harvard Dataverse, `https://doi.org/10.7910/DVN/OYHSJL`.

The folder includes the Stata code (Stata17) and data file (dta format) to reproduce the tables in the paper.

Educational attainment and Inter-generational Mobility: a Polygenic Score Analysis

Online Appendix

(Not meant to be part of the journal publication)

Aldo Rustichini, William Iacono, James Lee, Matt McGue

S-0.1. **Alternative Specifications of Parental Investment.** In this section we briefly outline the model in which investment of parents in human capital of children can affect human capital directly, but also the skill variable (θ).

The i^{th} household solves the optimization problem in the variables E expenditure in consumption, I^i pair of investment in the human capital and J^i of skill of the two children:

$$\max_{(E^i, J_1^i, J_2^i, I_1^i, I_2^i)} \mathbf{E}\left((1-\delta)\ln E^i + \delta \sum_{j=1,2} y_j^i\right), \tag{S-1}$$

subject to the budget constraint given by the household's income (y denotes the natural log of income):

$$E^i + \sum_{k=1,2} I_k^i + \sum_{k=1,2} J_k^i = \exp(y^i) \tag{S-2}$$

The expectation of equation (6) refers to the random shocks ϵ^h and ϵ^y.

The skill of twin ij is affected by a parental pecuniary investment J_j^i, in addition to the skill component in the genetic endowment, and is thus given by:

$$\theta_j^i = w(g_j^i) + \alpha_J \ln J_j^i + \Pi X_j^i + \epsilon_j^{\theta,i}. \tag{S-3}$$

The parameter α_J describes the effect of the parental investment on skill.

We assume the no-correlation and zero mean condition as in the main text. Human capital accumulation is described by equation 8, and income is given by equation 9, as in the main text. We assume zero mean for shocks to human capital and income as in the main text, and also assume that the shocks to human capital and income are not correlated.

At the optimal solution of the problem optimal parental investment is equal for the two siblings for both the component of the skill investment and the human capital investment, and is a constant fraction (depending on the parameters) of the total household income $exp(y^i)$, as in equation 4 of the main text. This equation can then be taken as the reduced form of the model presented in this section.

S-0.2. **MTAG correction of PGS.** We considered a different polygenic score using a correction that increases the predicitve power of the score. To do this, after the preminary stage indicated, we applied the software tool $MTAG$ Turley et al. (2019) to increase the effective sample size of the education summary statistics by drawing upon GWAS of IQ, a trait showing a strong genetic correlation with educational attainment. In this MTAG step, we used the IQ summary statistics of both Savage et al. (2018) and Lee et al. (2018). The weights of the SNP's in the score were then calculated with the software tool PRScs Ge et al. (2019), which uses an external sample to estimate the correlations between SNP's in order to convert the univariate regressions coefficients in $GWAS$ summary statistics to partial regression coefficients. PRScs also applies Bayesian shrinkage to the partial regression coefficients, which can then be used as weights in the polygenic score. We used the 1000 Genomes European populations to estimate the correlations between SNP's and calculated the shrunken partial regression coefficients of the 450,000 SNP's that were originally genotyped in MCTFR and survived all default software filters.

The two different scores yield very similar results in our analysis, hence which one we choose turns out to be of no substantial importance. We illustrate the difference in table S-1 below, that one can compare to the table 1 in the main text. The first column is identical, hence it is omitted. Similar comparisons are possible for the other estimates in the main text, with similar results.

The small size of the difference is probably due to the fact that the sample size of the underlying GWAS (Lee et al. (2018)) is large.

TABLE S-1. **Income at the age 29 take, family income, PGS, and Personality.** The PGS is MTAG-corrected. All variables, including College of parents and Male, are standardized to mean zero and SD 1. The signs of MPQ variables NA, Externalizing and Academic problems are reversed. Controlled for PC's and the parents-child time difference in age at income data collection.

	(1) b/se	(2) b/se
Family Income	0.127***	0.079**
	(0.027)	(0.032)
Male	0.276***	0.312***
	(0.025)	(0.029)
Male × Family Income	–0.061**	–0.050*
	(0.025)	(0.030)
PGS MTAG Corr	0.073***	0.006
	(0.025)	(0.029)
Education Years		0.257***
		(0.035)
IQ		0.011
		(0.029)
MPQ PA		0.061**
		(0.026)
MPQ NA		–0.024
		(0.027)
MPQ CN		0.034
		(0.032)
Externalizing		–0.072*
		(0.037)
Academic effort		0.057
		(0.038)
Academic problems		–0.017
		(0.034)
N	2100	1485

S-0.3. **Fixed Effects Analysis on *DZ* twins.** In this section we report results on fixed effects analysis on *DZ* twins, for Education Years, GPA, College and Intelligence.

TABLE S-2. **Education Years, only DZ twins.** All variables are standardized to mean zero and SD 1. Fixed effects regressions.

	(1) b/se	(2) b/se	(3) b/se
PGS Education	0.115*	0.114*	0.090
	(0.060)	(0.060)	(0.058)
MPQ PA		0.041	0.036
		(0.047)	(0.047)
MPQ NA		0.005	–0.028
		(0.050)	(0.049)
MPQ CN		0.071	–0.040
		(0.051)	(0.057)
Externalizing at 17			0.107
			(0.081)
Academic effort at 17			0.176**
			(0.076)
Academic problems at 17			0.029
			(0.064)
Constant	0.244***	0.273***	0.189***
	(0.027)	(0.040)	(0.054)
N	612	612	612

TABLE S-3. **GPA score: Fixed effects analysis in DZ twins.** All variables are standardized to mean zero and SD 1. Fixed effects regressions.

	(1) b/se	(2) b/se	(3) b/se
PGS	0.275***	0.179***	0.134***
	(0.055)	(0.053)	(0.044)
IQ		0.333***	0.186***
		(0.054)	(0.044)
MPQ PA		0.079*	0.063*
		(0.042)	(0.034)
MPQ NA		–0.001	–0.055
		(0.044)	(0.036)
MPQ CN		0.209***	0.002
		(0.045)	(0.042)
Externalizing at 17			0.106*
			(0.061)
Academic effort at 17			0.471***
			(0.057)
Academic problems at 17			0.102**
			(0.047)
Constant	–0.029	0.100***	–0.046
	(0.027)	(0.034)	(0.041)
N	682	630	590

TABLE S-4. **College and PGS in DZ twins: 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	2.851***	2.191***	1.904***
	(0.397)	(0.324)	(0.318)
IQ		3.507***	3.238***
		(0.637)	(0.670)
MPQ PA		1.291**	1.426**
		(0.166)	(0.216)
MPQ NA		1.345**	1.238
		(0.176)	(0.195)
MPQ CN		1.880***	1.075
		(0.270)	(0.193)
Externalizing at 17			1.517*
			(0.332)
Academic effort at 17			2.075***
			(0.480)
Academic problems at 17			1.350
			(0.258)
Constant	0.616***	1.008	0.713
	(0.086)	(0.160)	(0.153)
σ_u^2	3.898***	3.438***	3.423***
	(1.023)	(1.119)	(1.318)
N	865	780	645

TABLE S-5. **IQ score: Fixed effects analysis in DZ twins.** All variables are standardized to mean zero and SD 1. Fixed effects regressions.

	(1) b/se	(2) b/se	(3) b/se
PGS	0.152***	0.160***	0.125**
	(0.049)	(0.054)	(0.059)
MPQ PA		0.043	0.026
		(0.043)	(0.047)
MPQ NA		0.102**	0.108**
		(0.042)	(0.049)
MPQ CN		–0.089**	–0.161***
		(0.044)	(0.057)
Externalizing at 17			–0.151*
			(0.082)
Academic effort at 17			0.261***
			(0.075)
Academic problems at 17			0.097
			(0.064)
Constant	–0.069***	–0.058*	–0.054
	(0.024)	(0.034)	(0.056)
N	802	723	601

S-0.4. **Additional Evidence on Gene × Environment Correlation.** We report in this section regressions estimating the existence and effect size of gene × environment correlation. In each table the dependent variable of interested is regressed on the PGS of parents, and additional controls are considered. Both DZ and MZ twins are considered.

TABLE S-6. **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) b/se	(2) b/se	(3) b/se	(4) b/se
PGS	0.082***	0.097***	0.057*	0.069**
	(0.031)	(0.030)	(0.032)	(0.031)
PGS mother	0.105***	0.054**	0.064**	0.038
	(0.027)	(0.026)	(0.027)	(0.027)
PGS father	0.102***	0.020	0.036	–0.009
	(0.028)	(0.028)	(0.028)	(0.028)
IQ			0.152***	0.122***
			(0.023)	(0.023)
Soft Skills Index			0.222***	0.212***
			(0.022)	(0.022)
Education of parents		0.186***		0.116***
		(0.025)		(0.025)
Family Income		0.110***		0.083***
		(0.027)		(0.028)
Male	–0.091***	–0.081***	–0.039	–0.032
	(0.024)	(0.022)	(0.025)	(0.024)
Constant	0.291***	0.265***	0.317***	0.296***
	(0.023)	(0.023)	(0.023)	(0.023)
N	1686	1686	1333	1333

TABLE S-7. **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
PGS	0.217***	0.226***	0.120***	0.127***
	(0.033)	(0.033)	(0.031)	(0.031)
PGS mother	0.050	0.009	0.034	0.017
	(0.033)	(0.033)	(0.029)	(0.029)
PGS father	0.062*	–0.001	0.003	–0.025
	(0.034)	(0.035)	(0.030)	(0.031)
IQ			0.242***	0.226***
			(0.022)	(0.023)
Soft Skills Index			0.384***	0.380***
			(0.021)	(0.021)
Education of parents		0.158***		0.083***
		(0.032)		(0.028)
Family Income		0.078**		0.031
		(0.036)		(0.031)
Male	–0.226***	–0.221***	–0.136***	–0.132***
	(0.029)	(0.029)	(0.026)	(0.026)
Constant	0.024	0.001	0.038	0.026
	(0.029)	(0.029)	(0.025)	(0.025)
N	1579	1579	1389	1389

TABLE S-8. **College on PGS of Twin and PGS of parents, IQ and Soft Skills. Logit, Odds ratios displayed.** All independent variables are standardized to mean zero and SD 1. lnsig2u = panel level variance.

	(1) b/se	(2) b/se	(3) b/se	(4) b/se
PGS	1.817***	1.977***	1.616***	1.743***
	(0.276)	(0.291)	(0.285)	(0.303)
PGS mother	1.535***	1.158	1.306*	1.088
	(0.205)	(0.147)	(0.201)	(0.164)
PGS father	1.580***	1.051	1.138	0.873
	(0.221)	(0.142)	(0.181)	(0.139)
IQ			2.620***	2.193***
			(0.380)	(0.309)
Soft Skills Index			3.499***	3.326***
			(0.521)	(0.485)
Education of parents		2.607***		2.173***
		(0.333)		(0.323)
Family Income		1.612***		1.429**
		(0.216)		(0.229)
Male	0.648***	0.678***	0.838	0.867
	(0.075)	(0.073)	(0.116)	(0.116)
Constant	0.859	0.779**	0.996	0.905
	(0.098)	(0.085)	(0.130)	(0.116)
lnsig2u	6.407***	5.125***	6.314***	5.578***
	(1.076)	(0.903)	(1.322)	(1.212)
N	1805	1805	1411	1411

TABLE S-9. **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***	0.178***	0.173***	0.203***	0.187***
	(0.033)	(0.037)	(0.037)	(0.032)	(0.036)
PGS mother	0.068**	0.059*	0.053	0.000	–0.000
	(0.032)	(0.035)	(0.035)	(0.032)	(0.035)
PGS father	0.115***	0.122***	0.117***	0.032	0.046
	(0.032)	(0.036)	(0.036)	(0.033)	(0.037)
Soft Skills Index			0.087***		0.074***
			(0.024)		(0.024)
Education of parents				0.242***	0.221***
				(0.031)	(0.034)
Family Income				0.022	0.005
				(0.033)	(0.038)
Constant	–0.015	–0.009	–0.011	–0.033	–0.029
	(0.029)	(0.032)	(0.031)	(0.028)	(0.031)
N	1809	1415	1415	1809	1415

S-0.5. **Evidence of Genetic Assortative Mating.** Table S-10 shows the size of the genetic assortative mating, and that it is robust to control for possible population stratification, as the comparison between model (1), (2) and (3) confirms.

TABLE S-10. **PGS of parents.** Dependent variable: PGS of the mother. Model (3) controls for the square of each PC (not reported).

	(1) b/se	(2) b/se	(3) b/se
PGS of father	0.156***	0.133***	0.131***
	(0.033)	(0.034)	(0.034)
pc1		8.022***	5.024*
		(2.569)	(2.963)
pc2		–6.184**	–8.653***
		(2.657)	(2.885)
pc3		–2.094	–2.113
		(2.689)	(4.485)
pc4		–5.371**	–3.787
		(2.596)	(2.693)
pc5		1.239	2.545
		(2.687)	(2.721)
pc6		–1.634	–1.672
		(2.844)	(2.859)
pc7		4.146	3.311
		(2.779)	(2.793)
pc8		–4.025	–4.418
		(2.765)	(2.773)
pc9		–0.183	–0.496
		(2.866)	(2.867)
pc10		5.819**	5.826**
		(2.790)	(2.810)
N	951	918	918

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.

Barcellos, S. H., L. S. Carvalho, and P. Turley (2018): "Education can reduce health differences related to genetic risk of obesity," *Proceedings of the National Academy of Sciences*, 115, E9765–E9772.

Barth, D., N. W. Papageorge, and K. Thom (2020): "Genetic endowments and wealth inequality," *Journal of Political Economy*, 128, 1474–1522.

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.

——— (1989): "On the Economics of the Family: Reply to a Skeptic," *The American Economic Review*, 79, 514–518.

Becker, J., C. A. Burik, G. Goldman, N. Wang, H. Jayashankar, M. Bennett, D. W. Belsky, R. Karlsson Linnér, R. Ahlskog, A. Kleinman, et al. (2021): "Resource profile and user guide of the Polygenic Index Repository," *Nature human behaviour*, 5, 1744–1758.

Belsky, D. W., B. W. Domingue, R. Wedow, L. Arseneault, J. D. Boardman, A. Caspi, D. Conley, J. M. Fletcher, J. Freese, P. Herd, T. E. Moffitt, R. Poulton, K. Sicinski, J. Wertz, and K. M. Harris (2018): "Genetic analysis of social-class mobility in five longitudinal studies," *Proceedings of the National Academy of Sciences*, 115, E7275–E7284.

Björklund, A. and M. Jäntti (1997): "Intergenerational Income Mobility in Sweden Compared to the United States," *The American Economic Review*, 87, 1009–1018.

Björklund, A., J. Roine, and D. Waldenström (2012): "Intergenerational top income mobility in Sweden: Capitalist dynasties in the land of equal opportunity?" *Journal of Public Economics*, 96, 474 – 484.

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.

Black, S. E., P. J. Devereux, P. Lundborg, and K. Majlesi (2017): "On the Origins of Risk-Taking in Financial Markets," *The Journal of Finance*, 72, 2229–2278.

Blanden, J. (2011): "Cross-Country Rankings in Intergenerational Mobility: A Comparison of Approaches from Economics and Sociology," *Journal of Economic Surveys*, 27.

Bollen, K. A. (1989): *Structural Equations with Latent Variables*, New York u. a.: Wiley.

Cesarini, D. and P. M. Visscher (2017): "Genetics and educational attainment," *npj Science of Learning*, 2, 1–7.

Chiang, C., A. J. Scott, J. R. Davis, E. K. Tsang, X. Li, Y. Kim, T. Hadzic, F. N. Damani, L. Ganel, S. B. Montgomery, et al. (2017): "The impact of structural variation on human gene expression," *Nature genetics*, 49, 692–699.

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.

Ding, W., S. F. Lehrer, J. N. Rosenquist, and J. Audrain-McGovern (2009): "The impact of poor health on academic performance: New evidence using genetic markers," *Journal of health economics*, 28, 578—597.

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.

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.

Fletcher, J. and S. Lehrer (2011): "Genetic lotteries within families," *Journal of Health Economics*, 30, 647–659.

Galton, F. (1886): "Regression Towards Mediocrity in Hereditary Stature," *Anthropological Miscellanea*, 246–263.

Ge, T., C.-Y. Chen, Y. Ni, Y.-C. A. Feng, and J. W. Smoller (2019): "Polygenic prediction via Bayesian regression and continuous shrinkage priors," *Nature Communications*, 10, 1–10.

Goldberger, A. S. (1989): "Economic and Mechanical Models of Intergenerational Transmission," *The American Economic Review*, 79, 504–513.

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.

Heckman, J., R. Pinto, and P. Savelyev (2013): "Understanding the Mechanisms through Which an Influential Early Childhood Program Boosted Adult Outcomes," *American Economic Review*, 103, 2052–86.

Heckman, J. J. and T. Kautz (2012): "Hard evidence on soft skills," *Labour Economics*, 19, 451 – 464.

Himmelberg, C. (1972): "Fixed points of compact multifunctions," *Journal of Mathematical Analysis and Applications*, 38, 205 – 207.

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): "Gene–environment 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.

Kamae, T., U. Krengel, and G. L. O'Brien (1977): "Stochastic Inequalities on Partially Ordered Spaces," *Ann. Probab.*, 5, 899–912.

Knopik, V. S., J. M. Neiderhiser, J. C. DeFries, and R. Plomin (2017): *Behavioral genetics*, Worth Publishers, Macmillan Learning New York.

Kong, A., G. Thorleifsson, M. L. Frigge, B. J. Vilhjalmsson, A. I. Young, T. E. Thorgeirsson, S. Benonisdottir, A. Oddsson, B. V. Halldorsson, G. Masson, D. F. Gudbjartsson, A. Helgason, G. Bjornsdottir, U. Thorsteinsdottir, and K. Stefansson (2018): "The nature of nurture: Effects of parental genotypes," *Science*, 359, 424–428.

Lagakos, D., B. Moll, T. Porzio, N. Qian, and T. Schoellman (2018): "Life Cycle Wage Growth across Countries," *Journal of Political Economy*, 126, 797–849.

Lee, C.-I. and G. Solon (2009): "Trends in Intergenerational Income Mobility," *The Review of Economics and Statistics*, 91, 766–772.

Lee, J. 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.

Mazumder, B. (2005): "Fortunate Sons: New Estimates of Intergenerational Mobility in the United States Using Social Security Earnings Data," *The Review of Economics and Statistics*, 87, 235–255.

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.

McGue, M., A. Rustichini, and W. G. Iacono (2017): "Cognitive, noncognitive, and family background contributions to college attainment: A behavioral genetic perspective," *Journal of Personality*, 85, 65–78.

McGue, M., E. A. Willoughby, A. Rustichini, W. Johnson, W. G. Iacono, and J. J. Lee (2020): "The contribution of cognitive and noncognitive skills to intergenerational social mobility," *Psychological Science*, 31, 835–847.

Mincer, J. A. (1974): *Schooling, Experience, and Earnings*, no. minc74-1 in NBER Books, National Bureau of Economic Research, Inc.

Müller, A. and D. Stoyan (2002): *Comparison Methods for Stochastic Models and Risks*, Wiley.

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.

O'Brien, G. L. (1975): "The Comparison Method for Stochastic Processes," *Ann. Probab.*, 3, 80–88.

Okbay, A., Y. Wu, N. Wang, H. Jayashankar, M. Bennett, S. M. Nehzati, J. Sidorenko, H. Kweon, G. Goldman, T. Gjorgjieva, Y. Jiang, B. Hicks, C. Tian, D. A. Hinds, R. Ahlskog, P. K. E. Magnusson, S. Oskarsson, C. Hayward, A. Campbell, D. J. Porteous, J. Freese, P. Herd, 23andMe Research Team, Social Science Genetic Association Consortium, C. Watson, J. Jala, D. Conley, P. D. Koellinger, M. Johannesson, D. Laibson, M. N. Meyer, J. J. Lee, A. Kong, L. Yengo, D. Cesarini, P. Turley, P. M. Visscher, J. P. Beauchamp, D. J. Benjamin, and A. I. Young (2022): "Polygenic prediction of educational attainment within and between families from genome-wide association analyses in 3 million individuals," *Nature Genetics*, 54, 437–449.

Okbay, A. et al. (2016): "Genome-wide association study identifies 74 loci associated with educational attainment," *Nature*.

Österberg, T. (2000): "Inter-generational Income Mobility in Sweden: What do Tax-Data Show?" *Review of Income & Wealth*, 46, 421 – 436.

Palomino, J., G. Marrero, and J. G. Rodríguez (2018): "One size doesn't fit all: a quantile analysis of intergenerational income mobility in the U.S. (1980–2010)," *Journal of Economic Inequality*, 16, 347–367.

PARTHASARATHY, K. R. (1967): *Probability Measures on Metric Spaces*, Academic Press.

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.

ROGERS, A. R. (1983): "Assortative mating and the segregation variance," *Theoretical Population Biology*, 23, 110–113.

ROHLIN, V. A. (1952): "On the fundamental ideas of measure theory," in *American Mathematical Society Translation, n. 71.*

RUPERT, P. AND G. ZANELLA (2015): "Revisiting wage, earnings, and hours profiles," *Journal of Monetary Economics*, 72, 114–130.

RUSTICHINI, A., W. G. IACONO, AND M. MCGUE (2017): "The contribution of skills and family background to educational mobility," *The Scandinavian Journal of Economics*, 119, 148–177.

SATTLER, J. M. (1974): *Assessment of children's intelligence*, Saunders.

SAVAGE, J., P. JANSEN, S. STRINGER, K. WATANABE, J. BRYOIS, C. LEEUW, M. NAGEL, S. AWASTHI, P. BARR, J. COLEMAN, K. GRASBY, A. HAMMERSCHLAG, J. KAMINSKI, R. KARLSSON, E. KRAPOHL, M. LAM, M. NYGAARD, C. REYNOLDS, AND J. TRAMPUSH (2018): "Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence," *Nature Genetics*, 50.

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.

SOLON, G. (1992): "Intergenerational Income Mobility in the United States," *The 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.

Turley, P., R. Walters, and O. M. et al. (2019): "Multi-trait analysis of genome-wide association summary statistics using MTAG," *Nature Genetics*, 229–237.

Vilhjálmsson, B. J., J. Yang, H. K. Finucane, A. Gusev, S. Lindström, S. Ripke, G. Genovese, P.-R. Loh, G. Bhatia, R. Do, T. Hayeck, H.-H. Won, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Discovery, Biology, and Risk of Inherited Variants in Breast Cancer, S. Kathiresan, M. T. Pato, C. Pato, R. Tamimi, E. A. Stahl, N. A. Zaitlen, B. Pasaniuc, G. Belbin, E. E. Kenny, M. H. Schierup, P. L. De Jager, N. A. Patsopoulos, S. A. McCarroll, M. J. Daly, S. M. Purcell, D. I. Chasman, B. M. Neale, M. E. Goddard, P. M. Visscher, P. Kraft, N. Patterson, and A. L. Price (2015): "Modeling linkage disequilibrium increases the accuracy of polygenic risk scores," *American Journal of Human Genetics*, 97, 576–592.

Wang, B., J. R. Baldwin, T. Schoeler, R. Cheesman, W. Barkhuizen, F. Dudbridge, D. Bann, T. T. Morris, and J.-B. Pingault (2021): "Genetic nurture effects on education: a systematic review and meta-analysis," *bioRxiv*.

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 (2021): "The role of parental genotype in predicting offspring years of education: Evidence for genetic nurture," *Molecular Psychiatry*, 26, 3896–3904.

Yengo, L., A. R. Wood, S. Vedantam, E. Marouli, J. Sidorenko, S. Sakaue, S. Raghavan, G. Lettre, Y. Okada, J. N. Hirschhorn, and P. M. Visscher (2020): "A meta-analysis of height in 4.1 million European-ancestry individuals identifies ∼10,000 SNPs accounting for nearly all heritability attributable to common variants," Tech. rep., University of Queensland, Brisbane, Australia.

Young, A. I., S. M. Nehzati, S. Benonisdottir, A. Okbay, H. Jayashankar, C. Lee, D. Cesarini, D. J. Benjamin, P. Turley, and A. Kong (2022): "Mendelian imputation of parental genotypes improves estimates of direct genetic effects," *Nature Genetics*, 54, 897–905.

Zimmerman, D. J. (1992): "Regression Toward Mediocrity in Economic Stature," *The American Economic Review*, 82, 409–429.