# Trading social status for genetics in marriage markets: evidence from UK Biobank

Abdel Abdellaoui[*], Oana Borcan[†], Pierre Chiappori[‡] & David Hugh-Jones[§]

2022-04-11

**Abstract**

If socio-economic status (SES) and genetic variants are both assets in marriage markets, then the two will become associated in spouse pairs, and will be passed on together to future generations. This process provides a new explanation for the surprising persistence of inequality across generations, and for the genes-SES gradient (the genetic differences we observe between high- and low-income people). We model Social-Genetic Assortative Mating (SGAM) and test for its existence in a large genetically-informed survey. We compare spouses of individuals with different birth order, which is known to affect socio-economic status and which is exogenous to own genetic endowments among siblings. Spouses of earlier-born siblings have more genetic variants that predict educational attainment. We provide evidence that this effect is mediated by individuals' own educational attainment and income. Thus, environmental shocks to socio-economic status are reflected in the DNA of subsequent generations. Our work uncovers a new channel by which economic institutions can affect long-run inequality; suggests that genes-SES gradients may be historically widespread; and shows that genetic variation is endogenous to social institutions.

## Introduction

Over the long run, inequality is surprisingly persistent across generations (Clark and Cummins 2015; Solon 2018). Intergenerational mobility is correlated with cross-sectional inequality (Becker et al. 2018; Krueger 2012), which has risen dramatically in high-income countries, at the same time as intergenerational absolute mobility has declined (Western, Bloome, and Percheski 2008; Chetty et al. 2017).[1] Assortative mating in marriage markets can increase the

[*]Department of Psychiatry, Amsterdam UMC, University of Amsterdam, Amsterdam, The Netherlands. Email: a.abdellaoui@amsterdamumc.nl

[†]School of Economics, University of East Anglia, Norwich, UK. Email: O.Borcan@uea.ac.uk

[‡]Department of Economics, University of Columbia, New York. Email: pc2167@columbia.edu

[§]Corresponding author. School of Economics, University of East Anglia, Norwich, UK. Email: D.Hugh-Jones@uea.ac.uk

[1]Though relative mobility has been stable (Chetty et al. 2014). In the United Kingdom the Gini coefficient has increased from 26% to 34.6% between 1977 and 2020. The United States has seen a 10 percentage point rise to 43.3% during 1962-2013.

inequality of human capital and income across families (Breen and Salazar 2011; Greenwood et al. 2014). It follows that how families are formed, and transmit traits and assets to their offspring, are critical for understanding inequality. These processes have been studied from both socio-economic and genetic angles. While educational homogamy is well established, genetic assortative mating has been demonstrated only recently (Hugh-Jones et al. 2016; Robinson et al. 2017). Similarly, wealthy families pass on advantages to their children through both genetic inheritance and environmental influence (Rimfeld et al. 2018; Björklund, Lindahl, and Plug 2006).[2]

This paper examines a plausible, not previously analysed aspect of the spouse matching process: that both social status and genetics contribute to a person's attractiveness in marriage markets, and as a result, genetics and inherited social status may become associated in subsequent generations.[3] For example, suppose that wealth and intelligence are both positive assets in a potential spouse. Then wealthy people are more likely to marry intelligent people, and their children will inherit both wealth, and genetic variants associated with intelligence. We call this mechanism Social-Genetic Assortative Mating (SGAM). SGAM may be an important channel for the transmission of inequality. It leads to a hidden dimension of advantage for privileged families – hidden because most social science datasets do not include genetic information. This dimension may help to explain the surprising long-run persistence of inequality (Clark and Cummins 2015; Solon 2018). At the same time, this advantage is not an exogenous fact of biology, but endogenous to the social structure. Indeed, under SGAM, environmental shocks to an individual's social status may be reflected in the genetics of his or her children.

Below, we first outline a theoretical framework where attractiveness in the marriage market is a function of both socio-economic status (SES) and genetic variants. We show that social-genetic assortative mating in one generation increases the correlation between SES and genetic variants in the offspring generation. This result provides a new explanation of the *genes-SES gradient*, that is, the fact that SES is biologically heritable and associated with certain genetic variants (Belsky et al. 2018; Rimfeld et al. 2018; Björklund, Lindahl, and Plug 2006). The dominant existing explanation for the gradient is meritocratic social mobility: if a genetic variant predicts success in the labour market, then it will become associated with high SES and will be inherited in high-SES families. Under meritocracy, genes causes SES. On the other hand, under SGAM, causality goes both ways, from genes to SES and vice versa.

Next, using data on matched spouses born between 1935 and 1970 from UK Biobank (a large genetically-informed survey), we test the hypothesis that an individual's higher social status attracts spouses with higher genetic potential for educational attainment. Our genetic measure, the Polygenic Score for Educational Attainment (PSEA), derives from

[2] See Sacerdote (2011) for a review of the behavioural genetics and economics literatures on the nature vs nurture debate; for a broader review of intergenerational transmission of income see Black and Devereux (2010).

[3] *Social status* refers to characteristics that an individual possesses in virtue of their social position. For example, my wealth is a fact about me that holds in virtue of my relationship to certain social institutions (bank deposits, title deeds et cetera). Other examples include caste, class, income, and educational qualifications. *Socio-economic status* (SES) is a specific type of social status which exists in economically stratified societies, covering variables such as educational attainment, occupational class, income and wealth (e.g. White 1982).

large-scale genome-wide association studies (Lee et al. 2018) and is a cause of educational attainment itself, as well as of intelligence and labour market outcomes. It is already known that humans mate assortatively on PSEA, which makes it a likely candidate for detecting SGAM (Hugh-Jones et al. 2016; Robinson et al. 2017).

The endogeneity of socio-economic status is the main challenge in identifying the causal effect of SES on the spouse's genetic endowment. For instance, individuals with high education qualifications tend to also have high educational attainment genes, and as mentioned above, they may take partners based on genomic similarity. Indeed, recent studies show strong assortative mating on PSEA, much more than we would expect if spouses matched only on (our observed measures of) actual educational attainment (Okbay et al. 2022). To isolate the causal link from own SES to partner genes, we use the "accident of birth" as a shock to SES which is independent of own genetics. Specifically, we use an individual's birth order as a "treatment" which affects their partner choice through a range of mechanisms, including by affecting their own SES. It is well documented that earlier-born children enjoy higher parental investment and have better life outcomes, including measures of SES such as educational attainment and occupational status (Black, Devereux, and Salvanes 2011; Booth and Kee 2009; Lindahl 2008). At the same time, the facts of biology, in particular the so-called "lottery of meiosis", guarantee that siblings' birth order is independent of their genetic endowments.

Birth order could affect partner choice both through SES, and through non-SES mechanisms. So, we run a mediation analysis similar to Heckman, Pinto, and Savelyev (2013), decomposing the treatment effect into effects of measured and unmeasured mediating variables. Specifically, we estimate a reduced-form model with spouse polygenic scores for educational attainment (PSEA) as the dependent variable, and own birth order as the main independent variable. We then estimate a model which also includes measures of own socio-economic status, including university attendance and a measure of income. Under certain assumptions, these variables can be interpreted as mediating the effect of birth order on spouse genetics.

We find that later-born children have spouses with significantly lower PSEA in the reduced-form regressions. When we include university attendance and/or income as mediators, birth order is no longer independently significant, while the mediators increase the spouse's PSEA at 0.1% significance. University attendance explains an estimated 40-60% of the effect of birth order, and income explains about 10-13%. Thus, SES appears to mediate the effect of birth order on spouse genetics. The results are robust to the inclusion of several controls, including non-SES mediators, and a rich set of own genetic traits.

Our paper contributes to several literatures. Firstly, we study a new kind of assortative mating. The economics literature on matching in marriage markets has typically focused on educational similarities (e.g. Pencavel 1998; Chiappori, Salanié, and Weiss 2017) or social class or caste (e.g. Abramitzky, Delavande, and Vasconcelos 2011; Banerjee et al. 2013), but also sorting based on age, physical traits and ethnicity (Hitsch, Hortaçsu, and Ariely 2010). Matching

decisions on the marriage market follow multiple criteria, with some degree of substitutability between them.[4] For instance, Chiappori, Oreffice, and Quintana-Domeque (2012) showed that individuals trade off BMI for partners' income or education and that the marginal rate of substitution between these characteristics is different for males and females. The genetics literature has focused on genetic assortative mating (GAM), the phenomenon that people with similar genes marry each other. Recent research has confirmed the long-standing conjecture that GAM takes place in contemporary human populations (Howe et al. 2019; Hugh-Jones et al. 2016; Robinson et al. 2017). Geneticists have also developed the concept of cross-trait assortative mating (Beauchamp et al. 2010; Sundet et al. 2005). This happens when people with genes for e.g. height marrying people with genes for e.g. intelligence. As a result, the two types of genetic variation become associated. In this paper we bring the two literatures together, extending the idea of cross-trait assortative mating to encompass both social status, and genetic variants. Our results confirm that individuals with higher social status are more likely to attract a spouse with genetics for higher educational attainment.

Secondly, SGAM may affect economic inequality and intergenerational mobility. Clark and Cummins (2015) show using a database of surnames that long-run intergenerational persistence of wealth is higher than simple parent-child correlations would predict. Clark (2021) argues that the data can be explained by an underlying process where unobserved genetic variation determines wealth. We show below that SGAM could also generate these patterns. The mechanism again is unobserved genetic variation, but the interpretation is different, since we view genetic endowments not an exogenous source of variation, but as an asset effectively "traded" in marriage markets in exchange for wealth and social status. Put another way, for Clark, in analysing the effect of parental wealth on child wealth, genetics is a *confound*; under SGAM, it may also be a *mediator*.

SGAM also affects cross-sectional inequality, like other forms of assortative mating (Fernández and Rogerson 2001; Fernandez, Guner, and Knowles 2005; Eika, Mogstad, and Zafar 2019; Chiappori, Dias, and Meghir 2018). We know that there is a "genes-SES gradient": people with high and low SES have different genes. From twin studies, the heritability of occupational class and educational attainment, i.e. the proportion of variance explained by genetic differences between individuals, is around 50% (Tambs et al. 1989). Genome-wide Complex Trait Analysis shows that the family socio-economic status of 2-year-old children can be predicted from their genes (Trzaskowski et al. 2014). Children born into higher-income families have more genetic variants predicting educational attainment (Belsky et al. 2018). Studies comparing adoptees to non-adoptees show that both post-birth environment and pre-birth conditions (genetics and to a lesser extent prenatal environment) contribute to the transmission of wealth and human capital (e.g. Björklund, Lindahl, and Plug 2006). Thus, the genes-SES gradient is an important source of inequality. In particular, genetic variation in human capital is key, since a likely cause of the recent rise in inequality is the increase in market returns to human capital (e.g. Kaplan and Rauh 2013; Eika, Mogstad, and Zafar 2019).

[4] Oreffice and Quintana-Domeque (2010) show that height and BMI are associated with spouse earnings. Dupuy and Galichon (2014) find spouse matching on multiple independent dimensions, including education, height, BMI and personality.

SGAM shows how marriage markets can lead high SES to be associated with different genetic variants. Thus, it can explain the genes-SES gradient. The standard explanation for the genes-SES gradient is returns to talent in labour markets, a.k.a. meritocracy. Parents with higher ability reap higher market returns, and they may then pass both higher socio-economic status and their genes to their children, leading to an association between the two (Belsky et al. 2018). This mechanism depends on the level of meritocracy in social institutions (Branigan, McCallum, and Freese 2013; Heath et al. 1985): in a society where social status was ascribed rather than earned, it could not take effect. Indeed, after the fall of communism in Estonia, the heritability of SES increased, presumably because post-communist society allowed higher returns to talent (Rimfeld et al. 2018). By contrast, SGAM does not require meritocracy. Even when social status is entirely ascribed, it may still become associated with certain genetic variants, so long as their associated phenotypes are prized assets in marriage markets. Since meritocracy is historically rare, while assortative mating is universal, this suggests that genes-SES gradients are likely to be historically widespread.

Both meritocracy and SGAM may increase social inequality overall, if there are complementarities between genetic and environmental components of human capital, for example if higher-ability parents make more productive investments in children's human capital (Cunha and Heckman 2007; Cunha, Heckman, and Schennach 2010; Heckman and Mosso 2014; Kong et al. 2018), or if high-income parents are able to invest more in transmitting their human capital (Becker et al. 2018). Thus, by bringing "good genes" and enriched environments together, SGAM may increase inequality in the next generation.

Lastly, we contribute to a literature in economics that examines the relationship between genetic and economic variables. Benjamin et al. (2011) is an early review. Several more recent papers use polygenic scores, in particular polygenic scores for educational attainment (Barth, Papageorge, and Thom 2020; Papageorge and Thom 2020; Ronda et al. 2020). These papers – like the vast majority of the behavior genetics literature (see e.g. Plomin, DeFries, and McClearn 2008) – take genetic endowments as exogenous and examine how they affect individual outcomes, perhaps in interaction with the environment. We take a different approach by putting genetics on the "left hand side". Thus, our paper challenges the assumption, in economics and beyond, that genetic endowment is exogenous to economic characteristics. While this may be tenable in within-generation studies, it ceases to hold in intergenerational models. Social-genetic assortative mating is a causal mechanism going from socio-economic status to genetic traits.

Also, our model shows that the strength of this mechanism is endogenous to social and economic institutions. When SES is highly transmissible across the generations, this has the long-run effect of increasing the association between SES and genetics. If so, institutional reforms that increase *intergenerational mobility*, like mass education or inheritance taxation, may in the long run affect not only economic but genetic inequality. Conversely, an increase in *economic meritocracy* increases the association between SES and genetics in the long run.[5] This poses the problem raised first

[5] See Proposition **??** below.

by Young (1958), and more recently by Markovits (2019): meritocracy may be self-limiting or even self-undermining.

The observations behind SGAM are not new. That status and physical attractiveness assort in marriage markets is a commonplace and a perennial theme of literature. In the Iliad, powerful leaders fight over the beautiful slave-girl Bryseis. In Jane Austen's novels, wealth, attractiveness and "virtue" all make a good match. Marx (1844) wrote "the effect of ugliness, its repelling power, is destroyed by money." And Donald Trump claimed: "part of the beauty of me is that I am very rich." The literature on mate preference from evolutionary psychology (Buss and Barnes 1986; Buss 1989; Buss and Schmitt 2019) confirms that attractive mate characteristics include aspects of social status ("high earning capacity," "professional status") as well as traits that are partly under genetic influence ("intelligent," "tall," "kind," "physically attractive"). Despite this, we have found almost no previous work in genetics, economics or the social sciences that analyses SGAM or its consequences.[6]

## Model

People in the marriage market have two characteristics: $x = (x_1, x_2)$, drawn from a normal distribution

$$\mathcal{N}\left(\begin{array}{ccc} 0 & s^2 & \sigma \\ 0, & \sigma & S^2 \end{array}\right).$$

We interpret $x_1$ as a genetic measure, for example of genes predictive of height, physical attractiveness or intelligence. $x_2$ is a measure of socio-economic status, such as income or wealth, or social status more generally (we sometimes use "wealth" as a shorthand). The correlation between $x_1$ and $x_2$ is

$$Corr = \frac{\sigma}{sS} < 1.$$

People's attractiveness is given by

$$i\left(x\right) = ax_1 + \left(1 - a\right)x_2$$

where $a \in [0, 1]$ is a parameter reflecting the relative importance of genetics to wealth in the marriage market.[7] If $a = 0$, marriage markets are highly inegalitarian, such that only SES matters. If $a = 1$, marriage markets are egalitarian and only genetics matter. We expect realistic societies to fall between these extremes. Attractiveness $i$ is distributed

[6] Halsey (1958) showed in a two-class model that social mobility combined with assortative mating might increase the association between genetics and social class. Belsky et al. (2018) offer three reasons for the association between education-linked genetics and SES, but do not consider SGAM.

[7] Note that since the variance of the shocks to $x_1$ and $x_2$ (see below) has been normalized to 1, $a$ also reflects this variance. That is, a large variance of SES shocks (compared to genetic shocks) translates into $a$ being large.

$N(0, \sigma_I^2)$, where

$$\sigma_I^2 = a^2 s^2 + (1-a)^2 S^2 + 2a(1-a)\sigma.$$

People form matches with transferable utility, where the surplus for a match between $x$ and $y$ is $S(i(x), i(y))$ such that $\partial^2 S / \partial i \partial j > 0$, i.e. $S$ is supermodular. As a result there is positive assortative mating on attractiveness: $x$ matches with $y$ only if they are at the same quantile of attractiveness, i.e. if $i(x_1, x_2) = i(y_1, y_2)$. We describe this as social-genetic assortative mating (SGAM).

We also consider random matching as a benchmark to compare against SGAM. Under random matching, the distribution of couples' characteristics is normal with mean 0 and covariance matrix

$$\mathbb{C}\begin{pmatrix} x_1 \\ x_2 \\ y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} s^2 & \sigma & 0 & 0 \\ \sigma & S^2 & 0 & 0 \\ 0 & 0 & s^2 & \sigma \\ 0 & 0 & \sigma & S^2 \end{pmatrix}.$$

Our first proposition shows that SGAM leads to a positive correlation between one partner's wealth and the other's genetics.

**Proposition 1.** *Under SGAM, the distribution of couples' characteristics is normal, with mean 0 and covariance matrix*

$$\mathbb{C}\begin{pmatrix} x_1 \\ x_2 \\ y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} s^2 & \sigma & A^2 & AC \\ \sigma & S^2 & AC & C^2 \\ A^2 & AC & s^2 & \sigma \\ AC & C^2 & \sigma & S^2 \end{pmatrix} \tag{1}$$

*where:*

$$A = \frac{as^2 + (1-a)\sigma}{\sqrt{a^2 s^2 + (1-a)^2 S^2 + 2a(1-a)\sigma}} = \frac{as^2 + (1-a)\sigma}{\sigma_I};$$
$$C = \frac{a\sigma + (1-a)S^2}{\sqrt{a^2 s^2 + (1-a)^2 S^2 + 2a(1-a)\sigma}} = \frac{a\sigma + (1-a)S^2}{\sigma_I}.$$

*In particular, the covariance between $x_2$ and $y_1$, $AC$, is positive if either $\sigma > 0$ or $\sigma = 0$ and $0 < a < 1$.*

*Proof.* See Appendix. □

We consider the distribution of couples' wealth. Under random matching this has mean 0 and variance $2S^2$. Under SGAM, the variance is:

$$V(x_2 + y_2) = 2S^2 + 2C^2 > 2S^2$$

This is decreasing in $a$ and equals $4S^2$ if $a = 0$. Thus, SGAM increases cross-sectional inequality, but less so than pure matching on wealth.

## Children

All couples have the same number of children. A child's characteristics are given by:

$$\begin{aligned} x_1' &= \frac{\tau}{2}(x_1 + y_1) + \varepsilon \\ x_2' &= \frac{\theta}{2}(x_2 + y_2) + \eta \end{aligned} \tag{2}$$

where $x$ and $y$ are the child's parents, and $\varepsilon$ and $\eta$ are independent normal random shocks with mean 0 and variance 1.

Parameter $\tau \approx 1$ reflects genetic inheritance. Under standard biological assumptions $\tau = 1$ and characteristics show no regression to the mean. In our model this leads the variance of $x_1$ to grow without limit over generations. In reality, we expect $\tau < 1$ because very extreme characteristics are selected against, a process known as stabilizing selection (Schmalhausen 1949; Sanjak et al. 2018).

Parameter $\theta \in [0, 1]$ reflects inheritance of SES. Unlike $\tau$ it may vary between societies. $\theta$ is high when there is high intergenerational transmission of SES. Thus, $\theta$ captures social and economic institutions that affect this intergenerational transmission, from taxation and public education to hereditary nobility. If we interpret $x_2$ narrowly as wealth, $(1 - \theta)$ can be thought of as the rate of inheritance tax.

For the time being, we assume that a person's genetic endowment has no impact on their SES. Technically, thus, $x_2'$ does not directly depend on $x_1'$. In a meritocratic society we would expect adult SES to partly depend on genetics. We show that even absent meritocracy, correlations between $x_1'$ and $x_2'$ can arise.

We can now calculate the covariance matrix for $x' = (x_1', x_2')$ under SGAM as:

$$
\begin{aligned}
\mathbb{C} &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \begin{pmatrix} \frac{\tau}{2} & 0 & \frac{\tau}{2} & 0 \\ 0 & \frac{\theta}{2} & 0 & \frac{\theta}{2} \end{pmatrix} \begin{pmatrix} s^2 & \sigma & A^2 & AC \\ \sigma & S^2 & AC & C^2 \\ A^2 & AC & s^2 & \sigma \\ AC & C^2 & \sigma & S^2 \end{pmatrix} \begin{pmatrix} \frac{1}{2}\tau & 0 \\ 0 & \frac{1}{2}\theta \\ \frac{1}{2}\tau & 0 \\ 0 & \frac{1}{2}\theta \end{pmatrix} \\
&= \begin{pmatrix} \frac{1}{2}A^2\tau^2 + \frac{1}{2}s^2\tau^2 + 1 & \frac{1}{2}\theta\sigma\tau + \frac{1}{2}AC\theta\tau \\ \frac{1}{2}\theta\sigma\tau + \frac{1}{2}AC\theta\tau & \frac{1}{2}C^2\theta^2 + \frac{1}{2}S^2\theta^2 + 1 \end{pmatrix}
\end{aligned} \tag{3}
$$

We now explore two issues. First, under SGAM, genetic characteristics are no longer exogenous; because of assortative matching, they are (partly) socially determined. In particular, even if genetics and SES are uncorrelated among parents, the expected genetic endowment of the child is positively related to parental SES. Second, as a result, in the long run a correlation appears between traits; that is, high SES people have genes which are attractive in marriage markets.

Regarding point 1, we compute the expected genetic characteristic of the child, conditional on parental wealth:

$$
\mathbb{E}\left[\frac{\tau}{2}\left(x_1 + y_1\right) + \varepsilon \mid x_2 = v, y_2 = w\right]
$$

Given the symmetry of the model, this conditional expectation only depends on the parents' total wealth, i.e. $v + w$.

**Claim 1.** *Under random matching, the expected genetic endowment of the children is proportional to the parents' SES and to the covariance between SES and genetics for the parents. In particular, if $\sigma = 0$ (i.e. genetics and SES are uncorrelated for the parents), then the expected genetic endowment of the children does not depend on parental SES.*

**Claim 2.** *Under SGAM, if $\sigma = 0$ (i.e. genetics and SES are uncorrelated for the parents), then the expected genetic endowment of the children is linearly increasing in parental SES. The relationship increases with the ratio of genetic variance to SES variance, is zero for $a = 0$ or $a = 1$, and is highest for intermediate values of $a$.*

Next, we study the correlation between children's traits 1 and 2 as a function of $\sigma$, the covariance of parents' traits. We first consider the general case, then concentrate on $\sigma = 0$, i.e. when traits are initially uncorrelated.

**Claim 3.** *Under random matching, the correlation between characteristics is smaller for children than for parents. In particular, if genetics and SES are uncorrelated for the parents, then they are uncorrelated for the children.*

**Claim 4.** *Under SGAM, if genetics and SES are uncorrelated for the parents, then they are positively correlated for the children so long as $0 < a < 1$. The correlation is increasing in $\theta$.*

Whether characteristics are more or less correlated for children than for parents depends on whether the initial correlation between parents' characteristics is larger or smaller than the asymptotic one, derived below.

Figure **??** shows the intuition behind the model. Parents match on downward-sloping attractiveness isoquants given by $ax_1 + (1-a)x_2 = u$. Their children are in between them on both dimensions. This compresses the distribution along the attractiveness isoquants, which leads to a positive correlation between genetics and SES. The correlation between $x_1'$ and $x_2'$ is 0 when $a = 0$ or $a = 1$, because then spouses don't trade off SES for genes. It is highest for intermediate values of $a$.

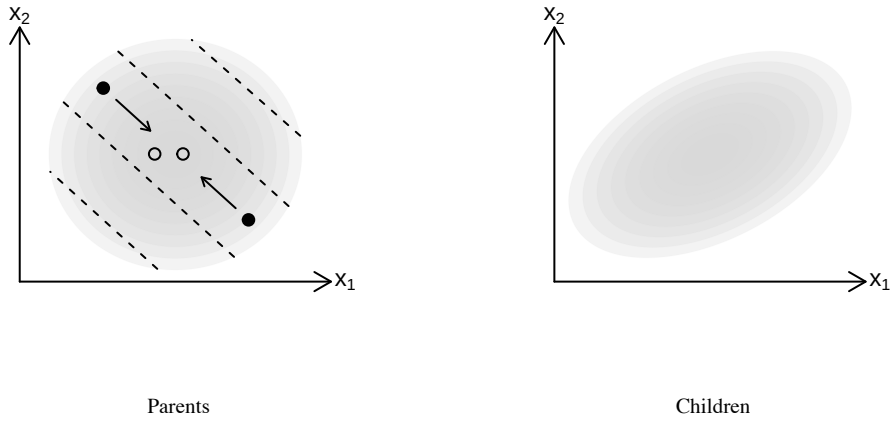


Figure 1: Theory. The shaded area is the population distribution. Parents (solid circles) match along attractiveness isoquants (dotted lines). Children (hollow circles) are between them. As a result, in the children's generation, the distribution is squeezed along attractiveness isoquants.

These results show that SGAM can lead to a genes-SES gradient, i.e. a positive correlation between genes and SES. Also, the strength of the genes-SES correlation is affected by economic institutions, as captured in $\theta$. When $\theta$ is high, the genes-SES correlation is high too.

We now consider the asymptotic distribution of $x_1$ and $x_2$ when the matching process is repeated over many generations. As we would expect, our main results continue to hold.

**Proposition 2.** *Under random matching, the dynamics converges to a stationary distribution that is normal with mean zero and covariance matrix*

$$\mathbb{C}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} \frac{2}{2-\tau^2} & 0 \\ 0 & \frac{2}{2-\theta^2} \end{pmatrix}$$

*In particular, the traits are asymptotically uncorrelated and children's expected genetic endowment is independent of parents' wealth.*

**Proposition 3.** *Under SGAM, for $\theta < 1$ and $\tau < 1$, the dynamics converge to a stationary distribution that is normal with mean zero and covariance matrix*

$$\mathbb{C}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} \bar{s}^2 & \bar{\sigma} \\ \bar{\sigma} & \bar{S}^2 \end{pmatrix}$$

*Moreover, the asymptotic correlation between characteristics, $corr = \bar{\sigma}/\bar{s}\bar{S}$, is non-negative, positive for $0 < a < 1$, increasing in $\theta$ and increasing then decreasing in $a$. The coefficient of parents' wealth on children's genetics is also positive for $0 < a < 1$.*

*For $\theta = 1$, the dynamics diverge and $\bar{S}^2$ goes to $+\infty$; for $\tau = 1$, the dynamics diverges and $\bar{s}^2$ goes to $+\infty$.*

Figure **??** plots the asymptotic correlation between $x_1$ and $x_2$ for $\tau = 0.95$.

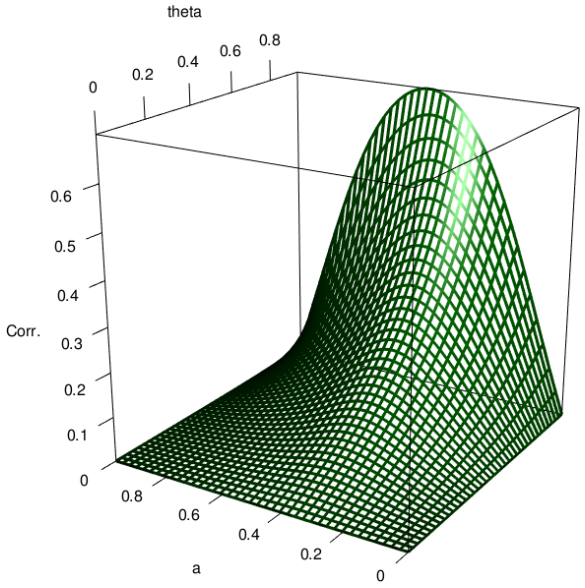


Figure 2: Long-run correlation between $x_1$ and $x_2$, by $a$ and $\theta$

Note that both $\bar{S}^2$ and $\bar{\sigma}$, as well as the correlation between characteristics and the conditional expectation of genetics given wealth, are increasing in $\theta$, i.e. decreasing in the tax rate. Higher taxation reduces the asymptotic variance of wealth (not surprisingly), but also the correlation between genetics and wealth.

## Extensions

We consider three extensions. First, the relative attractiveness of genes and SES might differ for men and women. Our basic result extends to this setup.

**Claim 5.** *Suppose that men's and women's attractiveness is given by*

$$i(x) = ax_1 + (1-a)x_2,$$
$$j(y) = by_1 + (1-b)y_2$$

*respectively, with* $0 \leq a \leq 1$, $0 \leq b \leq 1$. *Then if* $\sigma = 0$, *children's characteristics* $x_1'$ *and* $x_2'$ *will be positively correlated unless* $a = b = 0$ *or* $a = b = 1$. *The correlation is increasing in* $\theta$.

Interestingly, the $x_1$-$x_2$ correlation is highest when $a$ and $b$ are most different from each other. So gender differences in what counts as attractive make the effects of SGAM stronger.

Second, in modern meritocracies, people's adult SES depends not just on their parents' social status and on chance, but also on their own effort and skills, which might be related to their genetics. So, let

$$x_1' = \tau\frac{x_1 + y_1}{2} + \varepsilon$$
$$x_2' = \gamma x_1' + \theta\frac{x_2 + y_2}{2} + \eta \tag{4}$$

where $\gamma > 0$ represents the effect of own genetics on own SES. The basic result continues to hold, and also, the degree of meritocracy $\gamma$ increases the correlation between genes and SES; a highly meritocratic society may in the long run lead to a highly unfair genes-SES gradient.

**Proposition 4.** *Under SGAM and equation* (**??**)*, if genetics and SES are uncorrelated for the parents, then they are positively correlated for the children so long as* $0 < a < 1$ *or* $\gamma > 0$. *The correlation is increasing in* $\gamma$. *Also, so long as* $\gamma > 0$ *and either* $0 < a < 1$ *or* $\sigma > 0$, *the coefficient of parents' wealth on children's wealth exceeds* $\theta$.

Surprisingly, in this case, the children's correlation is not always increasing in $\theta$. The reason is that when $\gamma$ is high, a higher $\theta$ decreases the proportion of $x_2'$ that comes via $\gamma$ from own genetics, and increases the proportion that comes from parents' SES, which may be less strongly correlated with own genetics.

Third, we consider non-normal distributions of $x_1$ and $x_2$, non-normal shocks $\varepsilon$ and $\eta$, and non-linear attractiveness functions. Suppose

$$i(x) = f(ax_1, (1-a)x_2) \tag{5}$$

with $f$ strictly increasing in both its arguments. Our sole condition on the distribution of $x$ is that a positive measure of the population has attractiveness $i(x) = i$ where the distribution of $(x_1, x_2)|i$ is non-degenerate, i.e. not everybody with attractiveness $i$ is both genetically and socially identical. In particular, this allows for discrete distributions, like some kinds of social status.

**Proposition 5.** *Let attractiveness be given by* (**??**). *Let* $(x_1, x_2)$ *have any distribution such that a positive measure of the population has* $i(x) = i$ *where the conditional distribution of* $(x_1, x_2)|i(x) = i$ *is non-degenerate. Let* $\eta$ *and* $\varepsilon$ *be mean 0 and independent of* $x$ *and each other. If genetics and SES are uncorrelated for the parents, then the correlation among children is non-negative, and strictly positive if* $0 < a < 1$.

Other extensions are possible. We assumed that all couples have the same number of children. If fertility increased with $x_1$ or $x_2$, we would expect this to reduce the variance of traits in the children's generation and possibly also their covariance. Here, matching preferences, as summarized in the $a$ parameter, are exogenous. It would be natural to model $a$ as an equilibrium outcome. For example, if parents care about their children's wealth, $a$ might decrease in $\theta$ and increase in $\gamma$. Lastly, a gene-environment *interaction* (e.g. $x_2' = \alpha x_2 x_1$) might increase the gene-environment correlation some more.

## Discussion

The meanings of both social status, and "good genes" in the marriage market, are likely to vary across societies. Social status could encompass variables like social class or caste; ethnic identity in "ranked" ethnic systems; or in modern societies, SES, including wealth, income and occupation. Regarding genetics, standards of physical attractiveness, and other genetically-influenced characteristics which make someone a "good match", vary across societies and over time. By the logic of the model, in the long run, a positive correlation will emerge between any such characteristics and SES.

Recent empirical work shows high persistence of SES over time, in particular at the top. Clark (2021) argues that this is due to unobserved genetic variation. Proposition **??** shows that if genes affect own wealth directly, under assortative mating, the regression coefficient of parents' wealth on own wealth exceeds the "direct" coefficient $\theta$, because parents' wealth correlates with parents' genetics and via that with own wealth. Thus, regressions of wealth on wealth may include the effect of unobserved genetic variation. This may be a confound due to pre-existing gene-SES correlation (if $\sigma > 0$). But under SGAM it can also be a genuine cause, since changes in someone's wealth may indeed affect their spouse's genetics.

The converse also holds: regressions of children's characteristics on their genetics alone risk overestimating the effect of genetics, by confounding it with the effects of correlated socio-economic status. Recent work in genetics has shown this. Polygenic scores for educational attainment have smaller effects in between-sibling regressions, where between-family variation in SES is partialled out and where genetic variants are guaranteed to be randomly allocated, than in regressions which pool the whole sample (Howe et al. 2021). Parents' genetic variants which are *not* passed on to children predict children's characteristics, via environmental effects (Kong et al. 2018).

The model predicts variation in the strength of SGAM. In particular, in "caste societies" where there is complete

endogamy within social status groups, there is no scope for SGAM, because marriage partners do not trade off genetics for social status. Also, SGAM is increased by $\theta$. This implies that policy has long-run effects on the social structure: reducing $\theta$ not only increases intergenerational mobility, but reduces the correlation of genes with SES, and hence the unfairness of what has been called the "genetic lottery" (Harden 2021).

## Data and methods

In modern societies, both SGAM and meritocratic mobility may be at play. Genetic variants that cause higher SES, e.g. higher income and wealth, will be passed down along with that status. At the same time, higher SES and "good genes" will assort in the marriage market. To differentiate SGAM from meritocracy and from pre-existing correlations between genes and SES, we look for an environmental "shock" to people's SES, and examine its effects on the genetics of their spouse.

We use data from the UK Biobank, a study of about 500,000 individuals born between 1935 and 1970 (Bycroft et al. 2018). The Biobank contains information on respondents' genetics, derived from DNA microarrays, along with questionnaire data on health and social outcomes. The Biobank does not contain explicit information on spouse pairs. We categorize respondents as pairs if they:

- had the same home postcode on at least one occasion;[8]
- both reported the same homeownership/renting status, length of time at the address, and number of children;
- attended the same UK Biobank assessment center on the same day;
- both reported living with their spouse ("husband, wife or partner");
- consisted of one male and one female.

We also eliminate all pairs where either spouse appeared more than once in the data. This leaves a total of 35,682 pairs. Some of these could be false positives, i.e. people who are not each others' spouse but simply live in the same postcode. To validate the accuracy of our pairs, we use genetic relationships. Some respondents in the UK Biobank sample have a child (inferred from genetic data) who is also in the sample. Among our spouse pairs, 511 have a genetic child of at least one partner in the sample. For 441 of these, the child is the genetic child of both partners. If this subsample is representative, then at least 86% of the pairs who have had a child, have had a child together. This is a lower bound, because those who had a child with someone else may also have had a child with their partner who is not in the UK Biobank sample. As a point of comparison, 11% of families with dependent children included a stepchild in England and Wales in 2011 (National Statistics 2014).

It is still possible that some pairs in our data may not be actual spouses. In the appendix, to sign any possible bias

[8] A typical UK postcode contains about 15 properties.

in our estimates resulting from this, we use a dataset of "known fake" pairs. We show that estimated coefficients of interest are closer to zero among these fake pairs than among our candidate "real pairs". Because of this, any fake pairs remaining in our data are likely to bias our coefficients towards zero.

Our key dependent variable is spouse's *Polygenic Score for Educational Attainment* (PSEA). A polygenic score is a DNA-derived summary measure of genetic risk or propensity for a particular outcome, created from summing small effects of many common genetic variants, known as Single Nucleotide Polymorphisms (SNPs). We focus on PSEA rather than other polygenic scores for two reasons. First, educational attainment plays a key role in human mate search. People are attracted to educated potential partners (Buss and Barnes 1986; Belot and Francesconi 2013); spouse pairs often have similar levels of educational attainment, as well as similar PSEA (Vandenberg 1972; Schwartz and Mare 2005; Greenwood et al. 2014; Hugh-Jones et al. 2016). Second, PSEA predicts a set of important socioeconomic variables, including not only education but also social and geographic mobility, IQ, future income and wealth (Belsky et al. 2016; Barth, Papageorge, and Thom 2020; Papageorge and Thom 2020).[9]

We calculate PSEA using per-SNP summary statistics from Lee et al. (2018), re-estimated excluding UK Biobank participants.[10] We normalize the score to have mean 0 and variance 1. Because polygenic scores are created from estimates of many presumably tiny effects, they contain a large amount of noise relative to the true best estimator that could be derived from genetic data. For instance, PSEA explains only 11–13% of variance in educational attainment (out of sample, Lee et al. 2018), whereas the true proportion explained by genetic variation – the heritability – is estimated from twin studies to be about 40% (Branigan, McCallum, and Freese 2013). In addition, polygenic scores are no more guaranteed to be causal than any other independent variable. For example, social stratification by descent may lead genes to be associated with educational attainment even while playing no causal role (Selzam et al. 2019).

Despite these points, PSEA has non-trivial estimated effects on educational attainment. PSEA correlates with measures of education, including university attendance and years of full-time education; within-siblings regressions, where PSEA is randomly assigned by the "lottery of meiosis", confirm this correlation is at least partly causal (Lee et al. 2018). We recheck these facts within the UK Biobank sample. In a simple linear regression (N = 408,524) of university attendance on PSEA, a one-standard-deviation increase in PSEA was associated with a 9.2 percentage point increase in the probability of university attendance ($p < 2 \times 10^{-16}$). In a within-siblings regression among genetic full siblings (N = 36,748), the increase was 4.5 ($p < 2 \times 10^{-16}$). This suggests that about half of the raw correlation of PSEA with university attendance is down to confounds like good environments or parental nurture, while the remainder is causal. Still, the causal effect remains substantial: for a rough comparison, the (ITT) effect on college attendance of

[9] See Papageorge and Thom (2020) for a detailed discussion of polygenic scores, aimed at economists.

[10] PSEA was computed by summing the alleles across ~1.3 million genetic variants weighted by their effect sizes as estimated in genome-wide association studies (GWASs) that excluded UK Biobank. PSEA was then residualized on the first 100 principal components of the SNP array data. Further details can be found in Abdellaoui et al. (2019).

the Moving To Opportunity experiment in the US was 2.5 percentage points (Chetty, Hendren, and Katz 2016).

We use two measures of socio-economic status: income, and university attendance. Income is a direct measure of SES. University attendance is a predictor of income over the whole life course, and a form of SES in itself. The UK Biobank data only contains a direct measure of current household income, which is inappropriate for our purposes because it includes income from both spouses and is measured after marriage. Instead, we estimate income in £000s in the respondent's first job, by matching the job's Standard Occupational Classification (SOC) code with average earnings by SOC from National Statistics (2007). Job codes are only available for a subset of respondents.

Figure **??** illustrates the core idea of SGAM within our pair data. The X axis shows a measure of one partner's socio-economic status: university attendance or income. The Y axis plots the other partner's mean PSEA. Both males and females who went to university had spouses with higher PSEA. So did males and females with higher income. Since DNA is inherited, these people's children will also have higher PSEA.

These figures do not prove that SGAM is taking place. Since an individual's own PSEA correlates with both their educational attainment, and their income, both figures could be a result of genetic assortative mating (GAM) alone (Hugh-Jones et al. 2016). Indeed, recent studies show much higher levels of GAM than could be explained by matching on the observed education phenotype alone (Okbay et al. 2022). So, to demonstrate SGAM, we need a source of social status which is exogenous to genetics. Also, the link between social status and spouse genetics is likely to be noisy, for three reasons: first, polygenic scores contain a large amount of error, as discussed above; second, causal mechanisms behind variation in social status are likely to be noisy; third, to paraphrase Shakespeare (1595), the spouse matching process is highly unpredictable. So, we need a large N to give us sufficient power. This rules out time-limited shocks such as changes to the school leaving age (Davies et al. 2018).

We use *birth order*. It is known that earlier-born children receive more parental care and have better life outcomes, including measures of SES such as educational attainment and occupational status (Lindahl 2008; Booth and Kee 2009; Black, Devereux, and Salvanes 2011). On the other hand, all full siblings have the same *ex ante* expected genetic endowment from their parents, irrespective of their birth order. This is guaranteed by the biological mechanism of meiosis, which ensures that any gene is transmitted from either the mother or the father to the child, with independent 50% probability (Mendel 1865; Lawlor et al. 2008). For example, siblings' expected polygenic score is equal to the mean of their parents' polygenic scores.[11] We can therefore use birth order as a "shock" to social status. We put "shock" in quotes because we do not claim that birth order is exogenous to all other variables. For example, it naturally correlates with parental age, and it may also correlate with household SES at the time of birth. We only claim that birth order is exogenous to genetic variation.

[11] Although genetic variation is randomly assigned to children at birth, genetics and birth order could be dependent if parents' choice of whether to have more children is endogenous to the genetic endowment of their earlier children. We check for this below. Isungset et al. (2021) also find that birth order differences in education are not genetic.

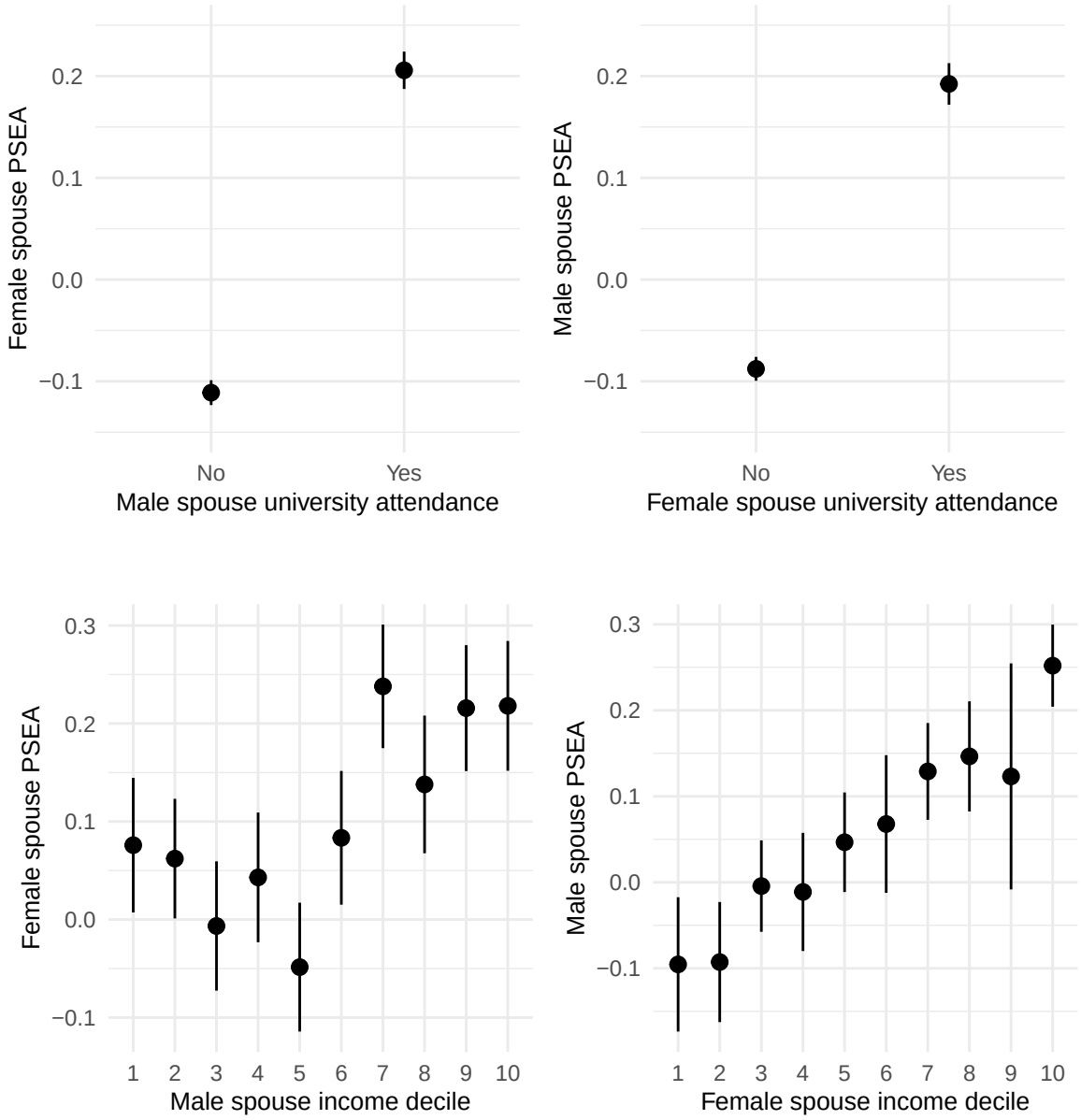


Figure 3: Spouse PSEA against own university attendance and own income in first job. Lines show 95% confidence intervals.

Our main independent variable is respondents' birth order, i.e. their number of elder siblings plus one. For controls we use family size, i.e. respondents' total number of siblings including themselves; month of birth; age at interview; respondents' own PSEA; and their father's and/or mother's age at their birth (calculated from parent's current age, only available if the parent was still alive). For most regressions, we use only respondents with between 1 and 5 siblings, i.e. with a family size of 2-6.

## Decomposing the birth order effect on spouse genetics

Ideally, we might prefer to use birth order as an instrument for SES. However, our measures of social status are noisy and incomplete. For example, we know whether subjects attended university, but not which university. Birth order likely affects both measured and unmeasured aspects of SES. So, an instrumental variables approach would be likely to fall foul of the exclusion restriction.

Instead, we conduct a mediation analysis, following the strategy of Heckman, Pinto, and Savelyev (2013). We first confirm statistically that birth order affects our measures of respondents' SES (income and education). Then, we regress spouse's PSEA on birth order, with and without controlling for SES. Under the assumption that birth order is exogenous to own genetics, these regressions identify the effect of birth order, plus other environmental variables that correlate with it, on own social status and spouse's genetics. Also, if the estimated effect of birth order on spouse's PSEA changes when SES is controlled for, that is evidence that SES mediates the effect of birth order.

We follow Heckman, Pinto, and Savelyev (2013) to decompose the aggregate treatment effect into components due to observed and unobserved proximate channels affected by the treatment. Our aim is to estimate the effect of SES (as an effect of birth order) on spouse PSEA.

Assume $B$ is a multivalued variable indicating birth order. Let $Y_b$ be the counterfactual outcome (spouse PSEA) for the first-born, second-born etc. Given $b$, spouse PSEA is assumed to be independent across observations conditional on some predetermined controls which are assumed not to be affected by $B$.

Let $m_b$ be a set of mediators, i.e. proximate outcomes determined by $b$, which account (at least in part) for the $b$ treatment effect on spouse PSEA. We can think of $m_b$ as all the effects on attractiveness, such as increments to SES, health, cognitive and non-cognitive skills, that individuals receive due to their birth rank. We can split the mediators in $m_b$ into a set $J_m$ of measured mediators, including university attendance and income in first job, and a set $J_u$ of mediators that we cannot measure.

Our linear model is:

$$Y_b = \kappa_b + \sum_{j \in J_m} \alpha_b^j m_b^j + \sum_{j \in J_u} \alpha_b^j m_b^j + \mathbf{X}'\boldsymbol{\beta_b} + \tilde{\varepsilon}_b = \tau_b + \sum_{j \in J_m} \alpha_b^j m_b^j + \mathbf{X}'\boldsymbol{\beta_b} + \varepsilon_b \tag{6}$$

where $\tilde{\varepsilon}_b$ is a mean-zero residual assumed independent of $m_b$ and $\mathbf{X}$; $\tau_b = \kappa_b + \sum_{j \in J_u} \alpha_b^j E(m_b^j)$; and $\varepsilon_b = \tilde{\varepsilon}_b + \sum_{j \in J_u} (m_b^j - E(m_b^j))$. We simplify by assuming that $\beta_b = \beta$ and $\alpha_b = \alpha$ for all $b$, i.e. that the effects of $\mathbf{X}$ and $m_B$ don't differ by birth order.[12] We assume differences in unmeasured investments due to $b$ are independent of $\mathbf{X}$.

We use a linear model for each observed mediator variable:

$$m_b^j = \mu_{0,j} + \mathbf{X}'\boldsymbol{\mu_{1,j}} + \mu_{2,j} \cdot b + \eta_j, j \in J_m \tag{7}$$

where $\eta_j$ is a mean-zero residual. We also assume the treatment-specific intercepts are linear in $b$:

$$\tau_b = \tau_0 + \tau b. \tag{8}$$

With the simplifying assumptions above and substituting (**??**) and (**??**) into (**??**) we obtain:

$$Y_b = \tau_0 + \tau b + \sum_{j \in J_m} \alpha^j (\mu_{0,j} + \mathbf{X}'\boldsymbol{\mu_{1,j}} + \mu_{2,j} \cdot b + \eta_j) + \mathbf{X}'\boldsymbol{\beta} + \varepsilon_b \tag{9}$$

Using equation (**??**), we can decompose the average treatment effect of a change from birth order $b$ to $b'$ into the effect of measured mediators $m^j$ and unmeasured mediators on the outcome:

$$E(Y_b' - Y_b) = \tau(b' - b) + \sum_{j \in J_m} \alpha^j E(m_{b'}^j - m_b^j) = \underbrace{\tau(b' - b)}_{\text{Effect of unmeasured mediators}} + \underbrace{\sum_{j \in J_m} \alpha^j \mu_{2,j}(b' - b)}_{\text{Effect of measured mediators}} \tag{10}$$

We are primarily interested in estimating the effect of SES on spouse PSEA, amongst the measured mediators, and furthermore we would like to measure the relative importance of SES compared to other factors in predicting spouse PSEA.

We therefore estimate:

$$Y = \tau_0 + \tau B + \sum_{j \in J_m} \alpha^j m_b^j + \mathbf{X}'\boldsymbol{\beta} + \varepsilon \tag{11}$$

[12] Under the assumption that measured and unmeasured mediators are uncorrelated, we can test these assumptions by running an OLS regression of an extended model (**??**) where we interact the measured mediators and controls with the treatment $B$, and test the significance of the coefficients on the interaction terms ($\boldsymbol{\alpha_b} = 0$ and $\boldsymbol{\beta_b} = 0$). See Heckman, Pinto, and Savelyev (2013) and Fagereng, Mogstad, and Rønning (2021) for details and different applications.

Estimating the above by OLS will generate unbiased estimates of $\alpha^j$ if $m^j$ is measured without error and is uncorrelated with the error term $\varepsilon$. Since $\varepsilon$ contains both individual disturbances and differences in unmeasured investments due to birth order, there are two identifying assumptions that need to hold for unbiased OLS estimates: (a) the measured investments (specifically SES) should be independent of unmeasured investments generated by birth order. Failing this, the estimates of $\alpha^j$ will be conflated with the effects of unmeasured investments. Second, (b) the measured investments should be uncorrelated with other shocks $\tilde{\varepsilon}_b$.

By running a least square regression of (**??**), we can estimate $\tau$ and $\alpha^j$. If assumption (a) holds, the part of the birth order treatment effect on spouse PSEA that is due to measured mediators, including SES, can be constructed using the estimated $\alpha^j$ and the effects of birth order on measured mediators. We can estimate these effects in a second step, from OLS regressions based on equation (**??**) for each measured mediator (in particular university attendance and income) on $\mathbf{X}$ and $B$. The part of the birth order effect that is due to university attendance (or income) on spouse PSEA will be the coefficient of university/income in the regression of spouse PSEA in equation (**??**), multiplied by the coefficient of birth order on university/income from equation (**??**).

# Results

We first regress our measures of socio-economic status, university attendance and income from first job, on birth order in our sample. We also do the same for four non-SES mediators that could be affected by birth order: fluid IQ, height, body mass index (BMI) and a measure of self-reported health. We control for respondent's own PSEA and their parents' age at birth (see below). Table **??** shows that birth order significantly predicts all these variables. Effect sizes are quite substantial: a single extra elder sibling reduces the chance of attending university by about 7.9 percentage points, income by about £1,089, fluid IQ by about 0.27 points on a 13 point test, height by about 0.7 centimeters, and self reported health by 0.043 points on a 4-point scale; and increases BMI by 0.19.

Next we run regressions of spouse PSEA on birth order, within our dataset of spouse pairs. Table **??** reports the results. Column 1 controls only for family size (using dummies). As expected, higher birth order is negatively associated with spouse's PSEA, though the estimated effect size is small and insignificant. Column 2 includes the respondent's own PSEA, as well as dummies for birth year to control for cohort effects, and dummies for birth month to control for seasonality effects. The effect size of birth order is not much changed.

Column 3 includes parents' age at birth. Within a family, later children have older parents by definition. Older parents have more life experience and may have higher income, which would presumably help later children.[13] Including

[13] We often only have data only for one parent. We use this, or take the mean if we have both. There are also potential genetic effects from parental age, though recent research has rejected these in favour of "social" explanations (Kristensen and Bjerkedal 2007; Black, Devereux, and Salvanes 2011). Cochran and Harpending (2013) report that mutational load is approximately linear in father's age, while it is constant in mother's age. We

Table 1: Regressions of variables on birth order

| | University | Income | Fluid IQ | Height | BMI | Health |
|---|---|---|---|---|---|---|
| Birth order | −0.0790 *** | −1.0899 * | −0.2733 *** | −0.7012 *** | 0.1907 ** | −0.0430 *** |
| | (0.0067) | (0.4264) | (0.0304) | (0.1355) | (0.0662) | (0.0103) |
| PSEA | 0.0889 *** | 1.5144 *** | 0.3180 *** | 0.1970 * | −0.4281 *** | 0.0533 *** |
| | (0.0046) | (0.3307) | (0.0200) | (0.0921) | (0.0456) | (0.0068) |
| Parents' age at birth | 0.0163 *** | 0.2623 *** | 0.0588 *** | 0.1514 *** | −0.0989 *** | 0.0110 *** |
| | (0.0012) | (0.0722) | (0.0053) | (0.0241) | (0.0117) | (0.0018) |
| Family size dummies | Yes | Yes | Yes | Yes | Yes | Yes |
| Birth month dummies | Yes | Yes | Yes | Yes | Yes | Yes |
| Birth year dummies | Yes | Yes | Yes | Yes | Yes | Yes |
| N | 10220 | 3412 | 10220 | 10220 | 10220 | 10220 |
| R2 | 0.074 | 0.026 | 0.058 | 0.017 | 0.023 | 0.018 |

Estimates from OLS regressions with the mediators (university attendance, income, fluid IQ, height, BMI, self-reported health) as dependent variables, and own birth order as the main independent variable. PSEA is the polygenic score for educational attainment, which is normalized with mean 0 and standard deviation 1. We include parents' age at birth (the mean of parents' ages) and further controls to ensure the balance of covariates across birth order. All data is from the UK Biobank for a sample of UK individuals born between 1935 and 1970. *** p $<$ 0.001; ** p $<$ 0.01; * p $<$ 0.05; + p $<$ 0.1. Standard errors: robust.

parents' age means we can separate the effect of parental age from birth order. This reduces the N by a lot, since only respondents with a live parent reported the necessary data. However, the effect of birth order jumps in size and becomes significant at the 5 per cent level. Meanwhile, parents' age has a positive effect. This suggests that estimates in columns 1-2 mixed two opposite-signed effects: having older parents versus being later in birth order.

Having tested that birth order affects spouse's PSEA, we now look for potential mediators of this effect. Despite the lower N, we continue to control for respondents' parents' age, since this removes a confound which would bias our results towards zero.[14]

Table **??** shows the results. Column 1 shows the effect of birth order, using the same specification as column 3 of the previous table. The remaining columns add potential mediators of birth order effects. Column 2 includes our first

observe very similar results if we control only for father's age at respondent's birth.

[14]The appendix reports results without controlling for parents' age.