Effects of inequality on genetics: evidence from UK Biobank

David Hugh-Jones & Abdel Abdellaoui

2020-08-05

1 Introduction

Charles Murray (1995) warned of "a merging of the cognitive elite with the affluent". On the opposite side of the political spectrum, Karl Marx (1844) wrote "I am ugly, but I can buy the most beautiful woman.... the effect of ugliness, its repelling power, is destroyed by money." These quotations suggest that social advantages, such as wealth, caste or status, may be transformed into biological advantages in the next generation, via assortative mating between socially and genetically advantaged people. We call this process genetic lock-in.

Figures 1 and 2 illustrate the idea using data for spouse pairs from UK Biobank. Figures 1 plots one partner's mean polygenic score for educational attainment (PSEA) against a measure of the other partner's actual educational attainment: possession of a university degree. University graduates had spouses with higher PSEA.[1] Figure 2 plots one partner's PSEA against another measure of social status: income.

These figures do not prove that genetic lock-in 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 partner selection on a purely genetic basis. In this paper, we test the theory more rigorously, using environmental shocks to social status that are unlikely to be correlated with own genetics. First, we develop a simple theory of genetic lock-in, to illustrate how its effects vary with social structure.

2 Theory

There is a large population, whose members have a single genetic trait g_i and a single social trait s_i.

Suppose that G and S are continuously distributed. Without loss of generality, $EG = ES = 0$.[2] People pair according to an attractiveness function

$$A(g_i, s_i) = f((1-k)g_i, ks_i)$$

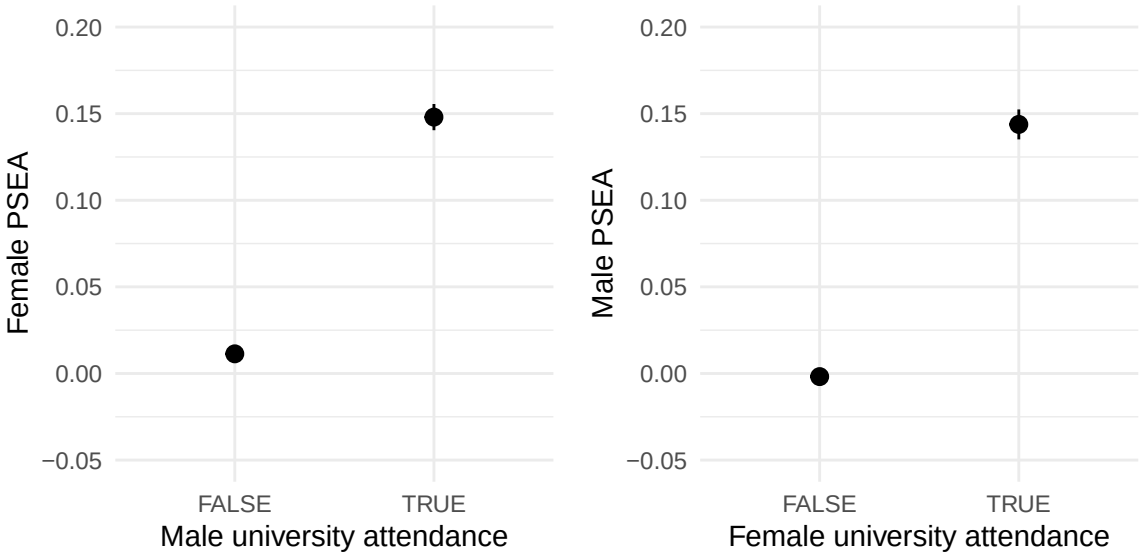
which is smooth and strictly increasing in both its arguments. If $k = 0$, "indifference curves" of attractiveness are vertical lines in (G, S) space. If $k = 1$, they are horizontal lines. If $k \in (0, 1)$ they are arbitrary downward sloping curves.

Write $p(i)$ for i's partner. Pairs always have the same attractiveness.

$$A(g_i, s_i) = A(g_{p(i)}, s_{p(i)}).$$

[1] To minimize concerns about genetic stratification, i.e. correlations between genetics and non-genetic forms of inherited advantage, PSEA is residualized by the first 100 principal components of UK Biobank array data.

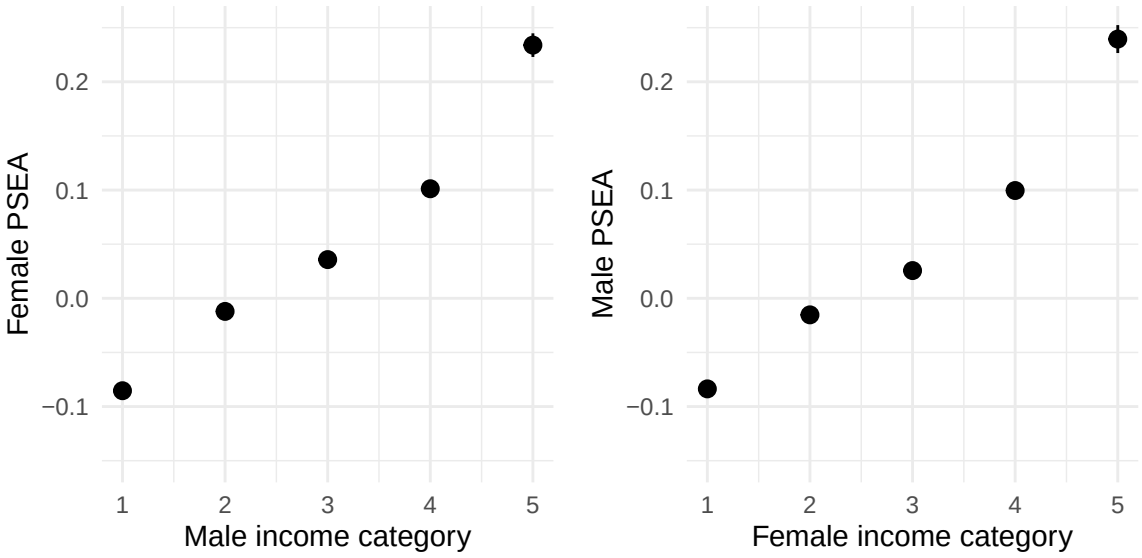
[2] Continuous distribution is not strictly required. All that is needed is for a set of pairs of positive measure to have different values of G and S, along a set of attractiveness curves of positive measure.



(a) Female PSEA by male educational attainment

(b) Male PSEA by female educational attainment

Figure 1: Social and genetic advantage among spouse pairs in UK Biobank



(a) Female EA3 by male income

(b) Male EA3 by female income

Figure 2: Social and genetic advantage among spouse pairs in UK Biobank

Each pair has two children. We assume that both children $c(d)$ of parents d, m have

$$\begin{aligned} g_{c(d)} &= \frac{g_d + g_m}{2}; \\ s_{c(d)} &= \frac{s_d + s_m}{2}. \end{aligned} \tag{1}$$

This is a strong assumption; we relax it later. For real world examples approximated by it, S could be wealth which is equally divided between the children; G could be a highly polygenic trait with many small effects. Write G_p, S_p to denote the population variables in the parents' generation; G_c, S_c for the children's generation.

Proposition 1. *(i) $Cov(G_c, S_c) \geq Cov(G_p, S_p)$, with strict inequality if and only if $0 < k < 1$.*

(ii) If $corr(G_p, S_p) \geq 0$, then $corr(G_c, S_c) \geq corr(G_p, S_p)$, with strict inequality if and only if $0 < k < 1$ or $corr(G_p, S_p) > 0$.

Proof. Within each pair $i, p(i)$ write d for the person with $s_d > s_{p(d)}$ and m for $p(d)$. (Think of these as "dukes" and "milkmaids", or if you prefer "duchesses" and "tennis instructors".) If $k < 1$, then $g_d < g_m$. (If $k = 0$, then define d as the person with $g_d < g_{p(d)}$.)

We integrate over the "dukes" to calculate the covariance in the parents' generation:

$$cov(G_p, S_p) = \int \frac{1}{2}(g_d s_d + g_{p(d)} s_{p(d)}) \, \mathrm{d}d.$$

For the children, the equivalent expression is

$$cov(G_c, S_c) = \int g_{c(d)} s_{c(d)} \mathrm{d}d,$$

observing that $EG_c = ES_c = 0$ from (1).

Take an arbitrary pair d, m. Write

$$\begin{aligned} g_d s_d &= (g_c - \Delta g)(s_c + \Delta s); \\ g_m s_m &= (g_c + \Delta g)(s_c - \Delta s) \end{aligned}$$

where

$$\begin{aligned} \Delta g &= \frac{g_m - g_d}{2} \geq 0, \text{strictly so if and only if } k > 0; \\ \Delta s &= \frac{s_d - s_m}{2} \geq 0 \text{strictly so if and only if } k < 1. \end{aligned}$$

Taking the average of the parents gives

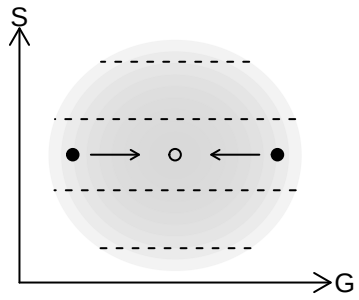
$$\frac{1}{2}(g_d s_d + g_m s_m) = g_c s_c - \Delta g \Delta s.$$

This is less than $g_c s_c$ if $0 < k < 1$, and equal to it if $k = 0$ or $k = 1$. Plugging this into the integral shows that

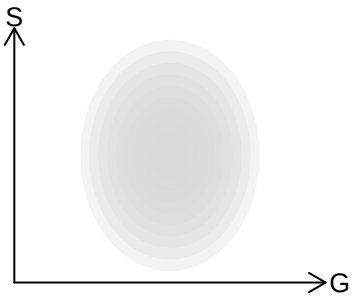
$$cov(G_p, S_p) \leq cov(G_c, S_c)$$

again with strict inequality if and only if $0 < k < 1$. This proves the first part. A similar argument, showing $var(G_c) \leq var(G_p)$ and $var(S_c) \leq var(S_p)$, proves the second part (see the appendix).

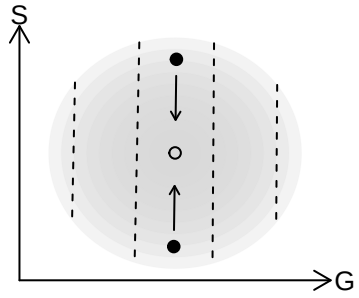
□



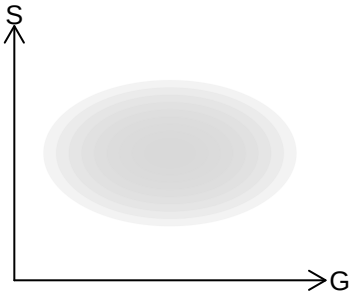
(a) Caste society (k = 1): parents



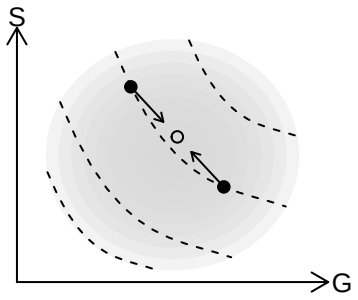
(b) Caste society (k = 1): children



(c) Egalitarian society (k = 0): parents



(d) Egalitarian society (k = 0): children



(e) Intermediate society (0 < k < 1): parents

Figure 3: Theory: shaded area is the population distribution. Dotted lines are attractiveness isoquants. Solid dots are example parents, transparent dots are example children. The right hand side shows the children's generation.

Figure 3 shows the intuition behind this result. The top row shows a caste society with $k = 1$. A typical pair is shown: children have intermediate values of G and S between their two parents (hollow circle). In this society pairs match only by social status; genetics plays no role. As a result, while the variance of G shrinks within each status group, genetics remain uncorrelated with social status in the children's population distribution, shown on the right. The next row shows an egalitarian society with $k = 0$. Parents match only by genetics and ignore social status. Again, as a result there is no correlation between genetics and social status in the children's generation. The bottom row shows an intermediate society. Because both genetics and social status contribute to attractiveness, matched spouses typically trade them off against each other. As a result, the distribution is squeezed along the gradient of k, and G and S are correlated in the children's generation.

Figure 6 in the Appendix shows that the condition in the second part cannot be relaxed further.

2.1 Robustness

We now relax the condition that children are exactly at the mean of their parents' values for G and S. Let

$$g_{c(i)} = \bar{g}_i + \varepsilon_i^G$$
$$s_{c(i)} = \bar{s}_i + \varepsilon_i^S$$

where

$$\bar{g}_i = \frac{g_i + g_{p(i)}}{2}; \bar{s}_i = \frac{s_i + s_{p(i)}}{2}$$

, ε^G has mean 0 and variance σ_G^2 and ε^S has mean 0 and variance σ_S^2.

Proposition 2. *1. if σ_G^2 and σ_S^2 are small enough and $corr(G_p, S_p) \geq 0$, then $corr(G_c, S_c) > corr(G_p, S_p)$ for $k \in (0, 1)$.*

2. if ε^G and ε^S are uncorrelated with each other and with $\bar{G}$ and $\bar{S}$; and if G_p and S_p are uncorrelated, then $corr(G_c, S_c) \geq 0$, with strict inequality if and only if $0 < k < 1$.

The conditions in Proposition are quite plausible. For G, they require that either variance in siblings' scores on some summary statistic is not too large, or that it is uncorrelated with the parents' scores. Both of these hold for most polygenic scores, which are additive sums of many small effects of alleles derived randomly from one or other parent. For S, the conditions would hold, for example, if S measures wealth, which is inherited not too unequally between siblings; or if wealth is inherited unequally but not in a way that correlates with S or G.

It is worth considering what kind of social arrangements would violate these conditions. For example, suppose that parents' combined wealth is inherited by the child with the lowest value of $g_{c(i)}$. This creates a negative correlation between $s_{c(i)}$ and $g_{c(i)}$.

2.2 Discussion

The "marriage market" here is a reduced form mechanism, encompassing that makes a difference to partner choice. For example, if earned income affects attractiveness in the marriage market, then society's level of meritocracy in the labour market will correlate with the value of k: a more meritocratic labour market will allow people with low social status but high human capital (partly genetically determined) to earn more, and therefore to enter the high group.

Also, the contents of G – what counts as "good genes" in the marriage market – are themselves likely to vary across societies. For instance, standards of physical attractiveness vary historically. Similarly, it is plausible that what counted as a "good match", in terms of personality, physical and intellectual characteristics, differed between medieval European nobility and contemporary society.

The model predicts variation in the strength of genetic lock-in. In particular, in "caste societies" where there is complete endogamy within social status groups, there is no scope for genetic lock-in, because marriage partners do not trade off genetics for social status. The model also assumes that social status is inherited randomly from one parent, in the same way a genetic allele is inherited. This assumption can be weakened. For example, if social status is inherited deterministically from the father, then the results remain unchanged (for each pair of parents, just assume that one randomly chosen parent is the father).

Behaviour geneticists often make the point that in meritocratic societies, successful people may transmit relevant genes to their offspring. (TODO: cite relevant papers.) Like genetic genetic lock-in, meritocracy may therefore lead to a correlation between social status and genetics. However, genetic lock-in is a distinct, though overlapping, mechanism. Under meritocracy, certain genetic variants cause higher social status and are then transmitted along with it. This logic does not apply in non-meritocratic societies where social status is ascribed rather than earned. Conversely, genetic variations which cause social status will become associated with it, even in the absence of assortative mating.

By contrast, genetic lock-in applies to genetic variants that are associated with higher social status in the spouse matching process. They do not need to exert any influence whatsoever on an individual's own social status. This process requires assortative mating, but does not require meritocracy. The logic of genetic lock-in therefore applies to a historically much wider range of societies, including societies where social status is wholly ascribed or inherited, such as aristocracies.

In modern societies, both assortative mating and meritocracy are likely to be at play. Genetic variants that cause (e.g.) higher income and wealth will be inherited along with components of social status such as inherited wealth, networks and cultural capital. At the same time, higher social status and "good genes" will assort in the marriage market, even if that higher social status is caused by purely environmental variation. Our empirical analysis shows this latter process at work.

3 Data

As mentioned above, simple correlations between one partner's social status and the other partner's genetics do not prove that genetic lock-in is taking place, because one's social status correlates with one's own genetics. To demonstrate genetic lock-in, we therefore need a source of social advantage which is exogenous to genetics. One possibility is birth order. It is well known that earlier-born children receive more parental care and have better life outcomes. (XXX is it? Go check.) On the other hand, early- and late-born full siblings have the same ex ante expected genetic endowment. [3] We can therefore use birth order as an exogenous shock to social status.

We use data from UK Biobank, a study of about 500,000 individuals.

- TODO: describe N for birth order, describe PSEA calculation.
- TODO: look at mechanisms by which birth order might affect university
- TODO: get IQ data, control for it
- TODO: subset to spouses with children
- TODO: overall index of social status?

[3] This might not be the case, if parents' choice of whether to have more children is endogenous to the genetic endowment of their earlier children. We will check for this below.

4 Results

Ideally we would instrument social status with birth order. However, our measures of social status are noisy and incomplete. For example, we know whether subjects went to university, but not which university they went to, and we only have rough categorical data on household income. Birth order likely affects both these and other measures of social advantage. So, an instrumental variables approach would probably fall foul of the exclusion restriction.

Instead, we conduct a mediation analysis, following the strategy of Heckman and Pinto (2013). We first regress our measures of own social status (i.e. income and education) on birth order. Then, we regress spouse's PSEA on birth order, with and without controlling for social status. 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 social status is included, that is evidence that social status mediates the effect of birth order.

- TODO: clarify the empirical model, you may need help….
- TODO: estimate individual income from job SIC codes? ASHE gives data

Figures 4 and 5 show the relationship between birth order, university education and income, separately for respondents with 1-3 siblings. We test this formally in a linear regression controlling for family size, which may be correlated with parents' characteristics including genetics. Birth order is negatively correlated with both measures (among respondents with 1-7 siblings: university p = 1.25e-23, income p = 4.27e-09).

```
## `summarise()` regrouping output by 'Birth order' (override with `.groups` argument)
```

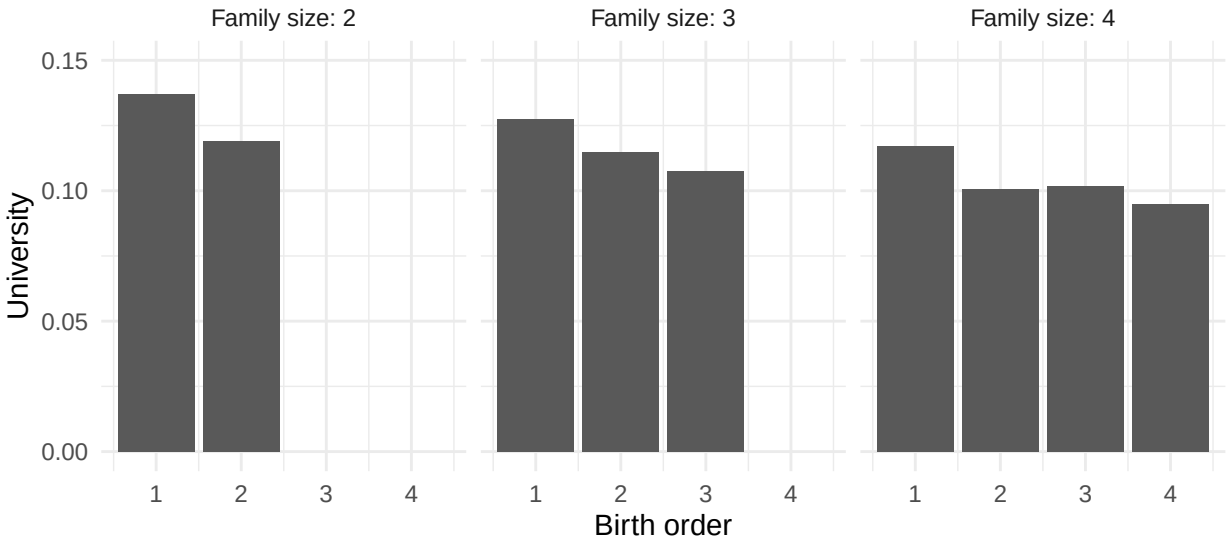


Figure 4: University attendance by birth order and family size

Next we run regressions of spouse PSEA on birth order, university attendance and income.

TODO: try to explain this better!

Table 1 shows the results. Column 1 shows the effect of birth order controlling only for own PSEA and family size. It establishes that earlier-born children have spouses with higher PSEA. The effect size is small. This is to be expected, because (a) the effects of birth order on university, income and (presumably) other variables

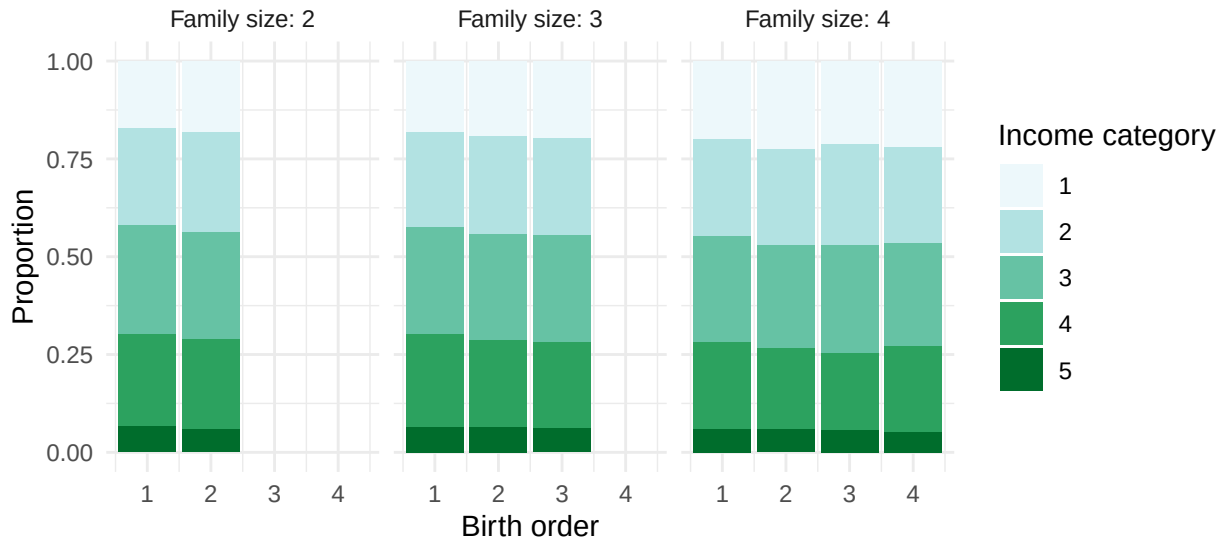


Figure 5: Income by birth order and family size

are small, and (b) PSEA is measured with a lot of error. We aim to test theory rather than estimating an effect size, so we focus more on statistical significance.

Column 2 includes university attendance. Column 3 includes income. Column 4 includes both. We estimate the percentage decrease in the effect of birth order across the columns, along with 95% confidence intervals for this figure, by running bootstraps (N = 199).[4] Including university attendance alone reduces the effect of birth order by 36.4% (CI 13.0% – 135.2%). Including income alone reduces the effect of birth order by 53.6% (CI 19.4% – 205.6%). Including both decreases the effect by 78.5% (CI 28.3% –297.0%).

4.1 Robustness

Although all children of the same parents have the same polygenic scores in expectation, it could still be possible that genetics correlates with birth order within the sample. This could happen if parents select family size on the basis of genetics. For example, if the first child had a phenotype reflecting a high (or low) polygenic score, then that might affect the parents' decision to have a second child. Alternatively, respondents might select into the sample on the basis of a combination of birth order and genetics. We check this by regressing 33 different polygenic scores on birth order, controlling for family size.[5] Table **??** shows the results. No scores were significant at $p < 0.10/33$. 3 scores were significant at $p < 0.10$ (body mass index, conscientousness, and neuroticism). Coefficients were never greater than 0.01 of a standard deviation. Table 2 in the appendix reruns regressions controlling for these scores. Results are almost unchanged.

5 Conclusion

TODO: Write!

[4]The sample percentage decrease calculated from the figures in Table 1 is not the correct estimate, since $E(X/Y) \neq EX/EY$.
[5]Polygenic scores were residualized on the first principal components of the genetic data.

Table 1: Regressions of spouse PSEA

	(1)	(2)	(3)	(4)
Birth order	-0.0050 *	-0.0037	-0.0031	-0.0023
	(0.0020)	(0.0020)	(0.0020)	(0.0020)
University		0.1103 ***		0.0837 ***
		(0.0054)		(0.0054)
Income			0.0627 ***	0.0595 ***
			(0.0016)	(0.0016)
Own EA3	0.0468 ***	0.0433 ***	0.0392 ***	0.0368 ***
	(0.0023)	(0.0023)	(0.0023)	(0.0023)
Family size dummies	Yes	Yes	Yes	Yes
N	301319	301319	301319	301319
R2	0.003	0.005	0.009	0.009
logLik	-427257.860	-427049.853	-426462.230	-426343.909
AIC	854533.721	854119.707	852944.460	852709.819

*** p < 0.001; ** p < 0.01; * p < 0.05. Standard errors clustered by spouse pair.

6 Appendix

6.1 Second part of Proposition 1

Proof. Write

$$corr(G_j, S_j) = \frac{cov(G_j, S_j)}{\sqrt{var(G_j)var(S_j)}} \text{ for both generations } j \in \{p, c\}. \tag{2}$$

where

$$var(G_p) = \frac{1}{2}\int g_d^2 + g_{p(d)}^2 \mathrm{d}d;$$
$$var(G_c) = \int g_{c(d)}^2 \mathrm{d}d.$$

Much as before,

$$\begin{aligned} g_d^2 + g_m^2 &= (g_c - \Delta g)^2 + (g_c + \Delta g)^2 \\ &= 2g_c^2 + 2(\Delta g)^2 \\ &\geq 2g_c^2. \end{aligned}$$

This shows that $var(G_c) \leq var(G_p)$ and a similar argument shows $var(S_c) \leq var(S_p)$. Thus the covariance is higher (and positive) in the children's generation, while the variances are lower. Combining these ensures that

$$corr(G_c, S_c) \geq corr(G_p, S_p).$$

Since for any k, either $var(G_c) < var(G_p)$ or $var(S_c) < var(S_p)$, the only way to get strict equality for the above is if $k \in \{0, 1\}$ and $cov(G_c, S_c) = cov(G_p, S_p) = 0$.

□

To show that the condition in the second part cannot be relaxed further, consider the distribution in Figure 6. There is negative correlation in the parents' generation (the shaded area). If $k = 1$ or is close enough to 1, then assortative mating along the dotted lines will reduce the variance of S along those lines, pushing the distribution towards the darker central area, without affecting the covariance. This will make the correlation more negative. After repeated generations the horizontal variance within values of G will almost disappear and the correlation will approach -1.

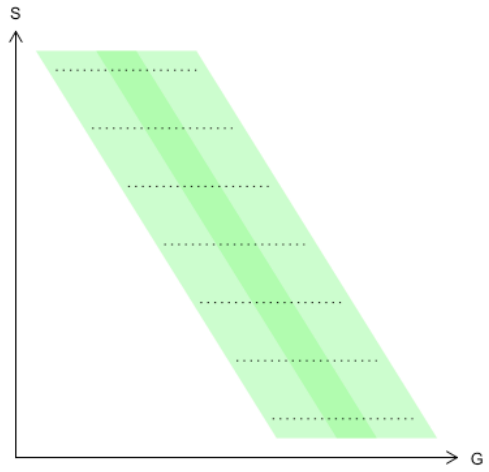


Figure 6: Correlation counterexample

6.2 Proposition 2.1

Proof. Note that in proposition 1, we took $g_{c(i)} = \bar{g}_i$ and $s_{c(i)} = \bar{s}_i$. Write

$$\begin{aligned} cov(G_c, S_c) &= cov(\bar{G} + \varepsilon^G, \bar{S} + \varepsilon^S) \\ &= cov(\bar{G}, \bar{S}) + cov(\varepsilon^G, \bar{S}) + cov(\bar{G}, \varepsilon^S) + cov(\varepsilon^G, \varepsilon^S). \end{aligned} \tag{3}$$

For any X and Y, $cov(X, Y)$ is bounded by $\sqrt{var(X)var(Y)}$. Plugging σ^2_G and σ^2_S into this formula shows that under condition 1, $cov(G_c, S_c)$ will be arbitrarily close to $cov(\bar{G}, \bar{S})$. Similarly, writing

$$var(G_c) = var(\bar{G}) + var(\varepsilon^G) + 2cov(\bar{G}, \varepsilon^G)$$

shows that $var(G_c)$ will approach $var(\bar{G})$ as σ^2_G grows small, and similarly for $var(S_c)$. Plugging these facts into (2) shows that $corr(G_c, S_c)$ approaches $corr(\bar{G}, \bar{S})$ as σ^2_G and σ^2_S grow small. Proposition 1 then shows $corr(\bar{G}, \bar{S}) < corr(G_p, S_p)$ for $k \in (0, 1)$.

Under condition 2, $cov(G_c, S_c) = cov(\bar{G}, \bar{S})$ since the last three terms of the sum in (3) are zero. Then since

$$cov(\bar{G}, \bar{S}) \geq cov(G_p, S_p) = 0$$

with strict inequality iff $k \in (0, 1)$, the covariance signs the correlation.

□

6.3 Regressions controlling for polygenic scores

Table 2: Regressions of spouse PSEA with controls for polygenic scores

	(1)	(2)	(3)	(4)
Birth order	-0.0050 *	-0.0038	-0.0032	-0.0023
	(0.0020)	(0.0020)	(0.0020)	(0.0020)
University		0.1100 ***		0.0835 ***
		(0.0054)		(0.0054)
Income			0.0625 ***	0.0594 ***
			(0.0016)	(0.0016)
Own EA3	0.0445 ***	0.0410 ***	0.0370 ***	0.0347 ***
	(0.0023)	(0.0023)	(0.0023)	(0.0023)
Family size dummies	Yes	Yes	Yes	Yes
Polygenic score controls	Yes	Yes	Yes	Yes
N	301319	301319	301319	301319
R2	0.004	0.005	0.009	0.010
logLik	-427236.323	-427029.346	-426444.468	-426326.693
AIC	854496.647	854084.692	852914.936	852681.386

*** p < 0.001; ** p < 0.01; * p < 0.05. Standard errors clustered by spouse pair.