

# Genetic lock-in of social advantage: evidence from UK Biobank

David Hugh-Jones & Abdel Abdellaoui

2020-08-04

## 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.<sup>1</sup> 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  $S_i$ . These are random variables which may be discrete or mixed and may also be correlated. Couples match and have children: we write  $G_p$  and  $S_p$  for the parents’ generation,  $G_c$  and  $S_c$  for the children’s generation, and drop the subscripts for statements which apply to both.

Couples match randomly among individuals of the same “attractiveness”, defined as

$$A = a(G, S) = (1 - k)f(G) + kv(S)$$

where  $f$  and  $v$  are continuous increasing functions. Without loss of generality, we measure  $G$  and  $S$  in “attractiles”, so that

$$A = (1 - k)G + kS.$$

We also scale them so that  $EG_p = ES_p = 0$ ,  $Var(G_p) = Var(S_p) = 1$ .

Define

$$B = kG - (1 - k)S.$$

---

<sup>1</sup>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.

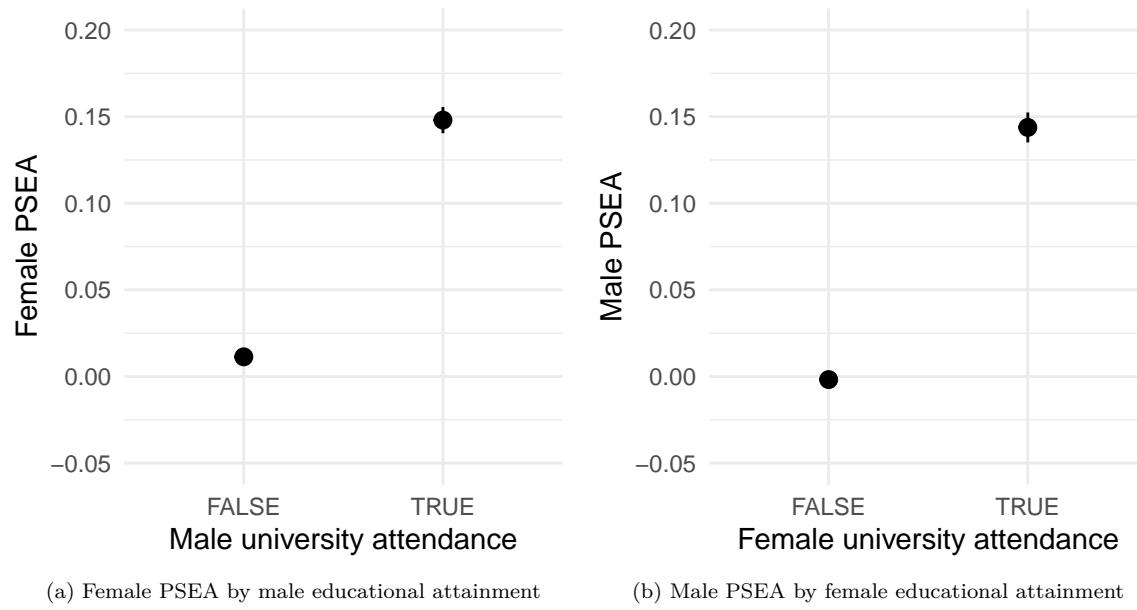


Figure 1: Social and genetic advantage among spouse pairs in UK Biobank

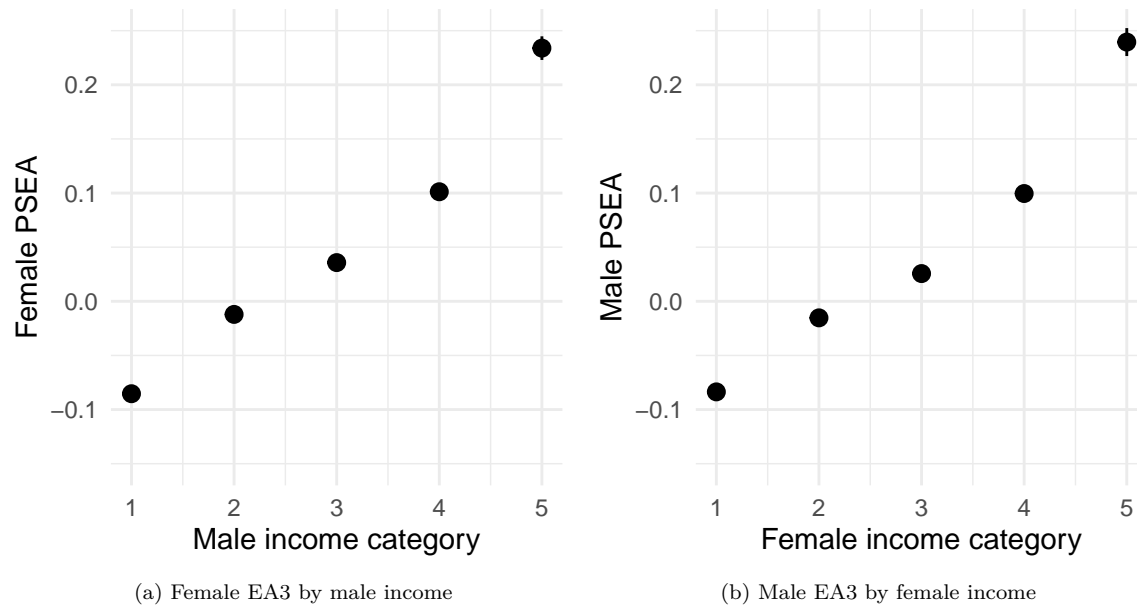


Figure 2: Social and genetic advantage among spouse pairs in UK Biobank

This is orthogonal to  $A$ . It can be thought of as the relative weight of “Beauty” or “Brains” in making someone attractive. We assume that

$$Var(B_p|A_p) > 0 \text{ for a set of values of } A_p \text{ of positive measure.} \quad (1)$$

Note that  $EA_p = EB_p = 0$ . Recall that if  $EX = EY = 0$ ,  $Var(X) = EX^2$  and  $Cov(X, Y) = EXY$ .

Couples have two children each, and children’s value of  $G$  and  $S$  is equal to the mean of their parents. As a result, children have the same value of  $A$  as both their parents.

Our main result that the correlation between genetic endowment and social status is higher in the children’s generation, so long as  $0 < k < 1$ .

**Proposition 1.** *When  $0 < k < 1$ ,  $Corr(G_c, S_c) > Corr(G_p, S_p)$ . When  $k = 0$  or  $k = 1$ ,  $Corr(G_c, S_c) = Corr(G_p, S_p)$ .*

Some points are in order. First, the proof does not depend strongly on the precise details of how  $S$  and  $G$  are inherited. All that is required is that the distribution of  $A$  does not change too much from parents to children, that along values of  $A$ ,  $E(B_c|A_p) \approx E(B_p|A_p)$ , and that

$$Var(B_c|A) \leq Var(B_p|A)$$

with strict inequality on a set of positive measure. These conditions would hold, for example, if  $S$  represented inherited wealth that was divided up between children not too unequally (and without taking account of the children’s  $G$ !); or if children’s expected value of  $G$  were equal to the mean of the parents’  $G$ , but with some variance.

Second, 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 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.<sup>2</sup> 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?

### 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 3 and 4 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  $p = 1.25\text{e-}23$ , income  $p = 4.27\text{e-}09$ ).

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

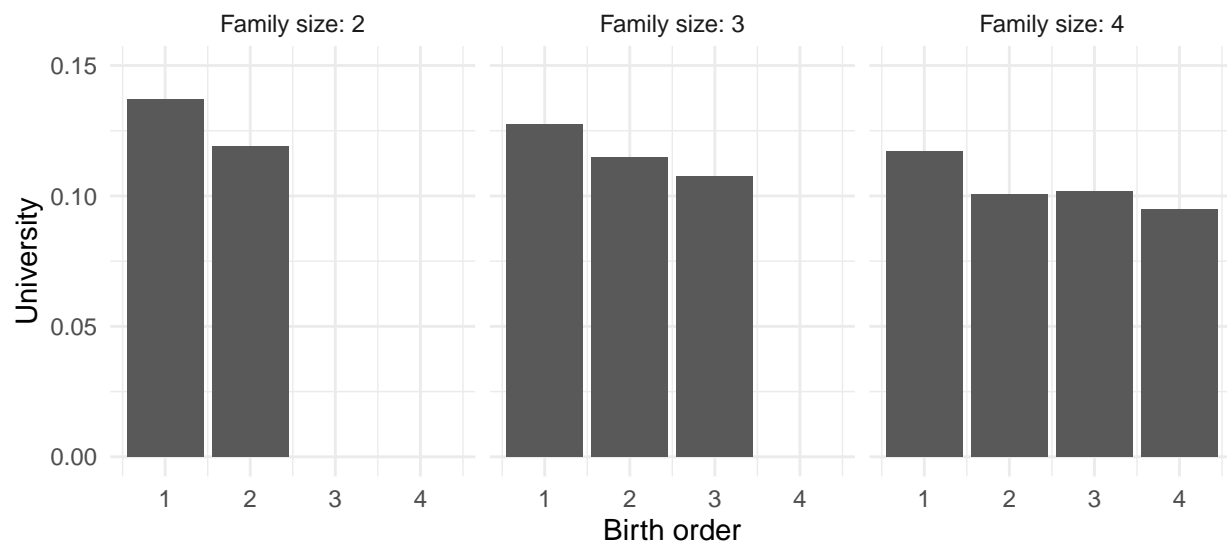


Figure 3: University attendance by birth order and family size

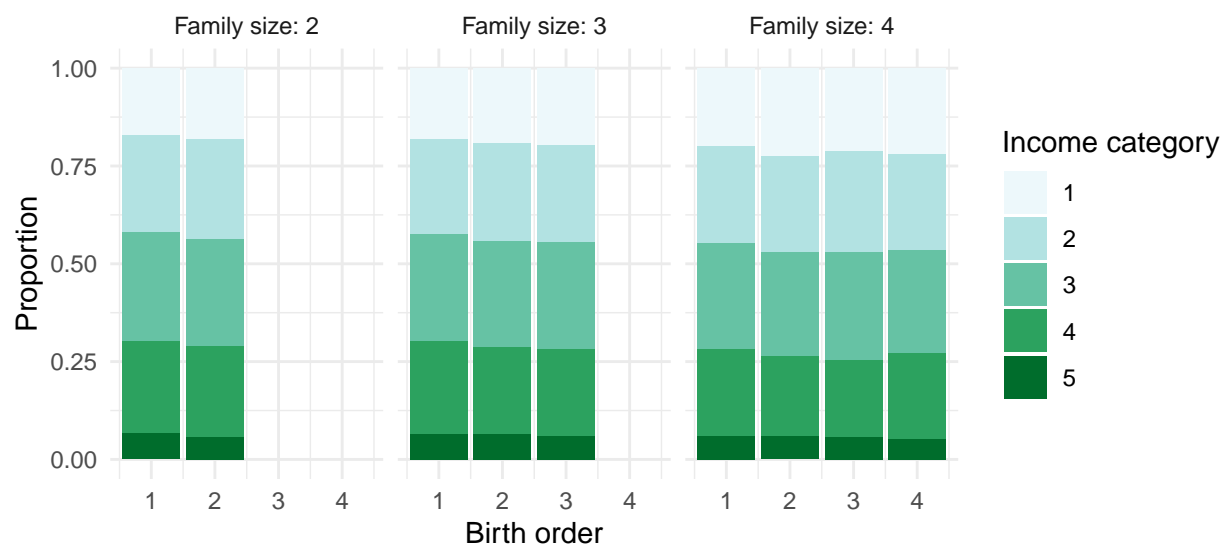


Figure 4: Income by birth order and family size

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

TODO: try to explain this better!

Table 1: 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.

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 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 ( $N = 199$ ).<sup>3</sup> 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

<sup>2</sup>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.

<sup>3</sup>The sample percentage decrease calculated from the figures in Table 1 is not the correct estimate, since  $E(X/Y) \neq EX/EY$ .

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.<sup>4</sup> Table ?? shows the results. No scores were significant at  $p < 0.10/33$ . 3 scores were significant at  $p < 0.10$  (body mass index, conscientiousness, 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!

## 6 Appendix

### 6.1 Proof of Proposition ??

Rearranging the equations gives

$$\begin{aligned} S &= cA - dB, \\ G &= dA + cB \end{aligned}$$

where

$$\begin{aligned} c &\equiv \frac{1-k}{z}, \\ d &\equiv \frac{k}{z}, \text{ and} \\ z &\equiv k^2 + (1-k)^2. \end{aligned}$$

We calculate the covariance of  $G$  and  $S$  in terms of the new variables:

$$\begin{aligned} Cov(G, S) &= E(GS) \\ &= E(dA + cB)(cA - dB) \\ &= E(cdA^2 + (c^2 - d^2)AB - cdB^2) \\ &= cd [Var(A) - Var(B)] + (c^2 - d^2)Cov(A, B) \end{aligned} \tag{2}$$

since  $EA = EB = EG = ES = 0$ .

Now we note that since children's  $A$  is equal to parents'  $A$ ,

$$Var(A_p) = Var(A_c), \tag{3}$$

and indeed the whole distribution is the same:

$$Pr(A_p = a) = Pr(A_c = a) \text{ for all } a. \tag{4}$$

Next, note that for every value of  $A$ ,

$$E(B_c|A) = E(B_p|A), \tag{5}$$

since within each pair, the children's value of  $B$  is equal to the parents' mean.

Lastly, for every value of  $A$ , the children's  $B$  is the mean of two random draws from  $B$  given  $A$ . Therefore:

$$Var(B_c|A) = \frac{1}{2}Var(B_p|A). \tag{6}$$

---

<sup>4</sup>Polygenic scores were residualized on the first principal components of the genetic data.

Applying the law of total variance to  $B$ , and using (6) and (5), along with (1):

$$\begin{aligned} \text{Var}(B_c) &= E_A(\text{Var}(B_c|A)) + \text{Var}(E(B_c|A)) \\ &< E_A(\text{Var}(B_p|A)) + \text{Var}(E(B_p|A)) = \text{Var}(B_p). \end{aligned} \tag{7}$$

Rewrite

$$\begin{aligned} &\text{Cov}(A, B) \\ &= \int \int ab \, \text{Pr}(a)\text{Pr}(b|a) \, db \, da \\ &= \int a \, \text{Pr}(a) \int b\text{Pr}(b|a) \, db \, da \\ &= \int a \, \text{Pr}(a)E(B|A) \, da. \end{aligned}$$

Using (4) and (5) then shows that

$$\text{Cov}(A_p, B_p) = \text{Cov}(A_c, B_c). \tag{8}$$

Plugging facts(8), (7) and (8) into the expression in (2), we have shown

$$\begin{aligned} \text{Cov}(G_p, S_p) &= cd[\text{Var}(A_p) - \text{Var}(B_p)] + (c^2 - d^2)\text{Cov}(A_p, B_p) \\ &\leq cd[\text{Var}(A_c) - \text{Var}(B_c)] + (c^2 - d^2)\text{Cov}(A_c, B_c) \\ &= \text{Cov}(G_c, S_c). \end{aligned}$$

The inequality is strict when  $cd$  is positive, equivalently when  $k \in (0, 1)$ , and holds with equality otherwise. The covariance and correlation share the same sign, so this completes the proof.

## 6.2 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.