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

David Hugh-Jones & Abdel Abdellaoui

2020-07-31

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

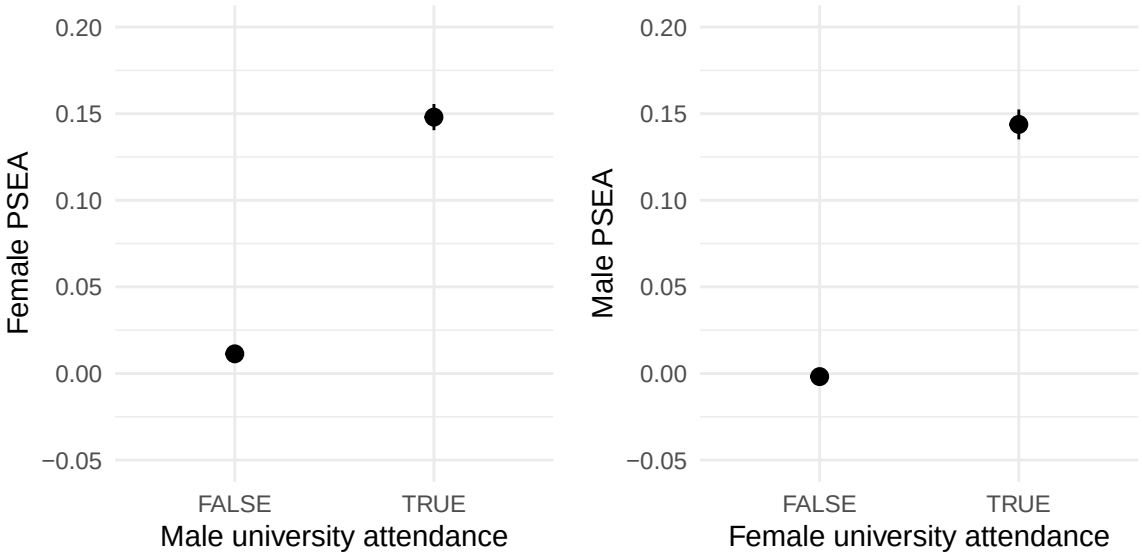
The population is size 1. There is a single binary genetic trait and a single social trait. Individuals i thus have $G_i \in \{0,1\}$ and $S_i \in \{0,1\}$; write $\tau_i = (G_i, S_i)$ to fully describe i's type. We will call individuals with $\tau_i = (1,1)$, $(1,0)$, $(0,1)$ and $(0,0)$, b-types, g-types, s-types and n-types respectively. The proportion of the population with $G_i = 1$ is denoted by γ, and the population proportion with $S_i = 1$ is denoted by σ. In generation 1, G_i and S_i are assumed independent, so e.g. the proportion of b-types is $\gamma\sigma$.

Higher values of G_i and S_i are assumed to be attractive in the marriage market. Indeed, individuals are divided into two classes, H and L, for the purpose of marital matching. Within H and L, all individuals form spouse pairs at random. Class H is of size θ: we assume $\gamma\sigma < \theta \leq \min\{\gamma, \sigma\}$. We also assume γ and σ are not too different:

$$\text{Either } \sigma < \gamma < (1+\sigma)/2 \text{ or } \gamma < \sigma < (1+\gamma)/2. \tag{1}$$

Married pairs have two children and each randomly inherits one parent's genetic trait and one parent's social trait. As a result, σ and γ stay constant over generations.

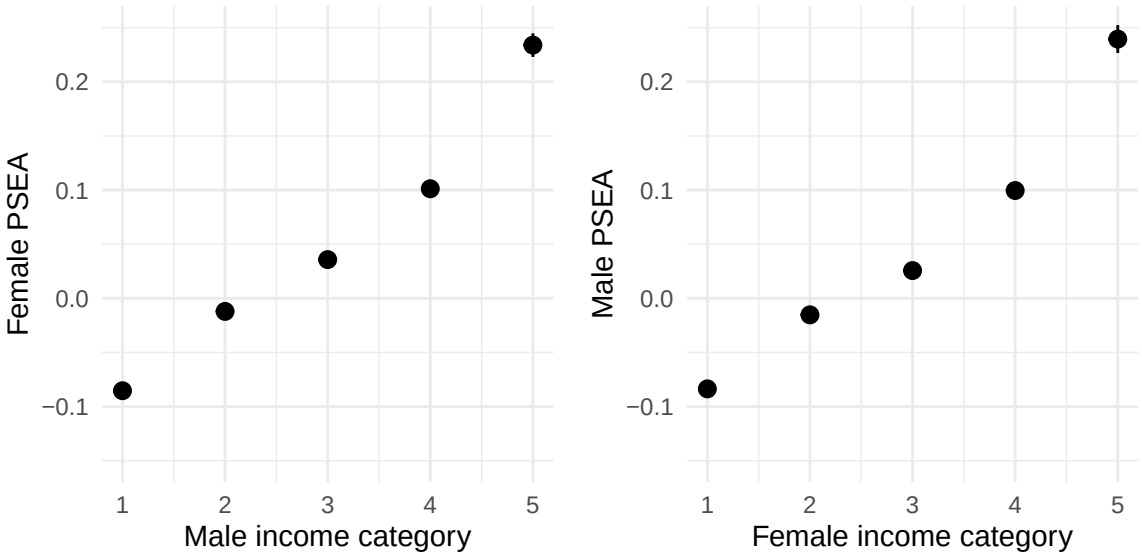
[1] To minimize concerns about genetic stratification, i.e. correlations between genetics and non-genetic forms of inherited advantage, PSEA is residualized by the first 100 principal components of UK Biobank array data.



(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



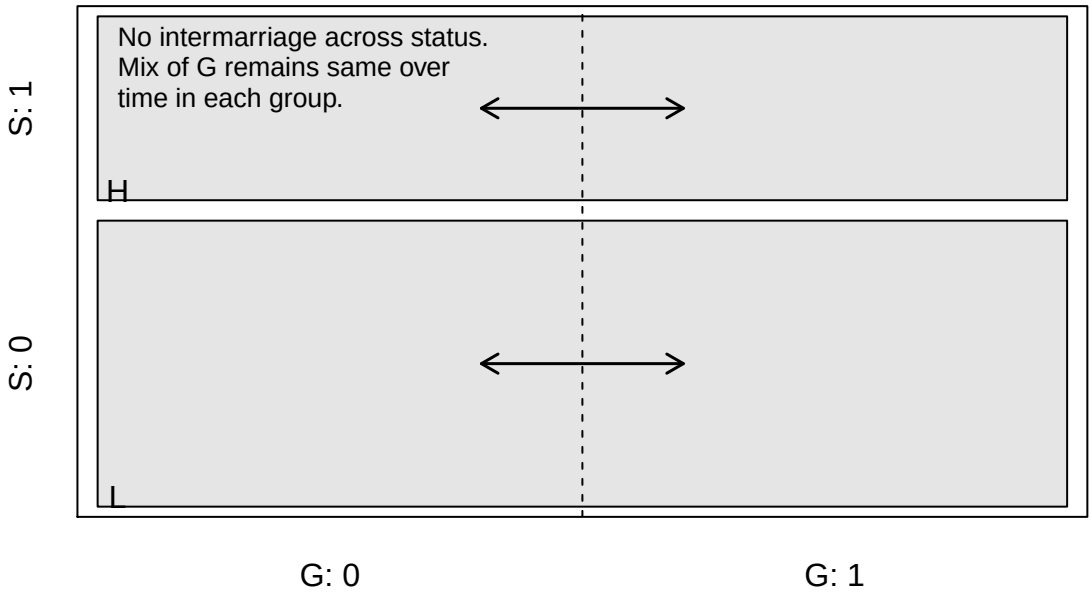
(a) Female EA3 by male income

(b) Male EA3 by female income

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

B-types with $\tau_i = (1,1)$ are always in H. N-types with $\tau_i = (0,0)$ are always in L. Where the other groups fit depends on the nature of the society. Of the $\theta - \gamma\sigma$ non-b-types in H, a proportion k are s-types and the remaining $1-k$ are g-types.

The value of k gives the level of social stratification in the society's marriage market. When k is close to 1, individuals' place in the marriage market is largely determined by social status, and genetics alone rarely gives access to the high matching group. When k is close to 0, "good genes" play a larger role in the marriage market, and social status alone rarely gives access to the high group.



(a) Caste society

Figure 3: Genetic lock-in: example societies

To build intuition, we start with two extreme examples. Suppose first that $k = 1$ and that $\theta = \sigma$. This is a caste society where marital matching is entirely driven by social status. All individuals in H (L) have $S_i = 1$ ($S_i = 0$), and so do their children. On the other hand, the proportion of individuals with $G_i = 1$ stays at γ within both groups. Thus, over generations, the distributions of G_i and S_i remain independent in the population.

Suppose next that $k = 0$ and $\theta = \gamma$. This society is socially egalitarian: only genetics matter for marital matching. All individuals in H (L) have $G_i = 1$ ($G_i = 0$), and so do their children. The proportion of individuals with $S_i = 1$ stays at σ within both groups; hence, again, G_i and S_i remain independent over generations.

In between these cases, G_i and S_i will become positively correlated. We examine the correlation in the second generation.

Proposition 1. *The second-generation correlation between G_i and S_i is concave in k. If θ is close enough to $\min\{\sigma, \gamma\}$, then the correlation has a maximum at $k \in (0,1)$.*

The logic behind Proposition 1 is simple. As we have seen, when spouse matching is entirely on the basis of social status, or entirely on the basis of genetics, no correlation arises between the two. However, when spouses are selected on the basis of both genetics and social status, then some s-types, who only have high social status, marry some g-types, who only have valuable genetic capital. As a result, some of their children inherit both (or neither), and this increases the correlation between genetics and social status.

Figure 4 plots the second generation correlation coefficient of S_i and G_i against k, for $\gamma = \sigma = \theta = 0.5$.

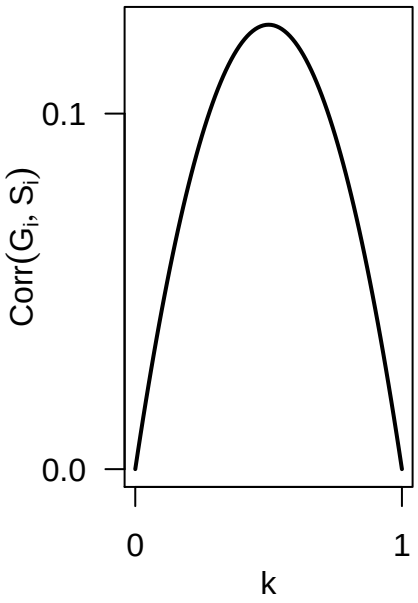


Figure 4: Social stratification and genetic lock-in: theory

Some points are in order. First, 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.

Second, 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).

Another possibility is that cultural rules may assign status more or less strictly. Suppose that if both parents have $S_i = 1$, then the children always inherit $S_i = 1$. However, if only one parent has $S_i = 1$, then children inherit $S_i = 1$ with probability α. A higher value of α corresponds to a more "expansive" inheritance rule for social status. For example, while British noble titles descend only to the first born son, on Europe the status of nobility was granted to all children, creating a separate caste. (TODO: check details!) The original

model corresponds to $\alpha = \frac{1}{2}$.[2]

Proposition 2. *The second-generation correlation between G_i and S_i decreases in α.*

TODO: Points about the theory.

- A more complex model with continuous S and G would still have the same results (maybe discuss in appendix)?

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?

[2] If $\alpha \neq \frac{1}{2}$, the size of the high-status group increases or decreases over time. Population shares could be stabilized by adding noise to the inheritance process, however.

[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 5 and 6 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)
```

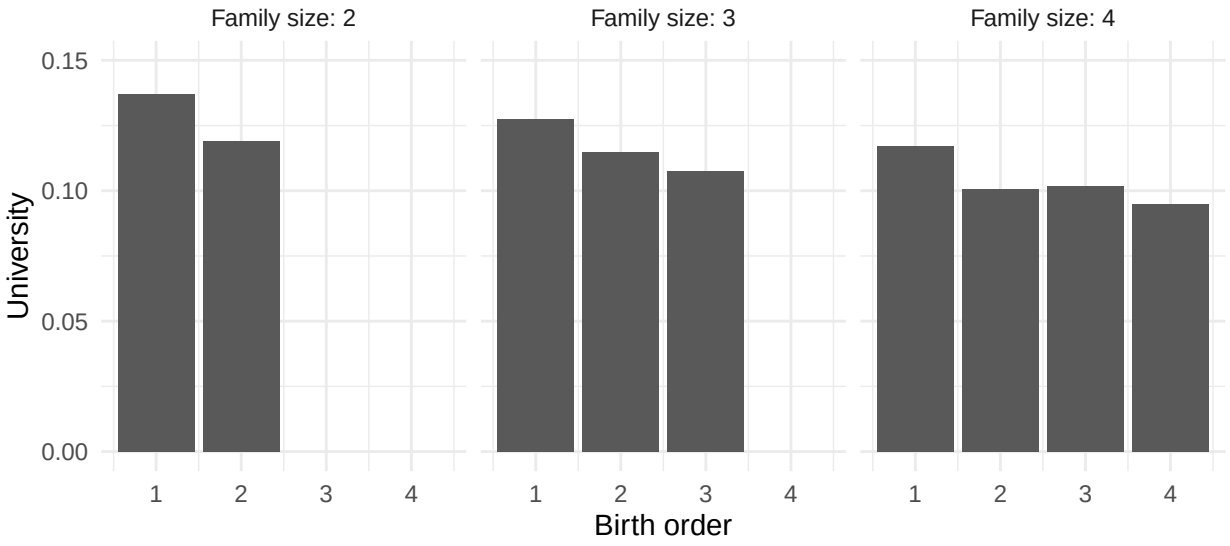


Figure 5: University attendance by birth order and family size

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

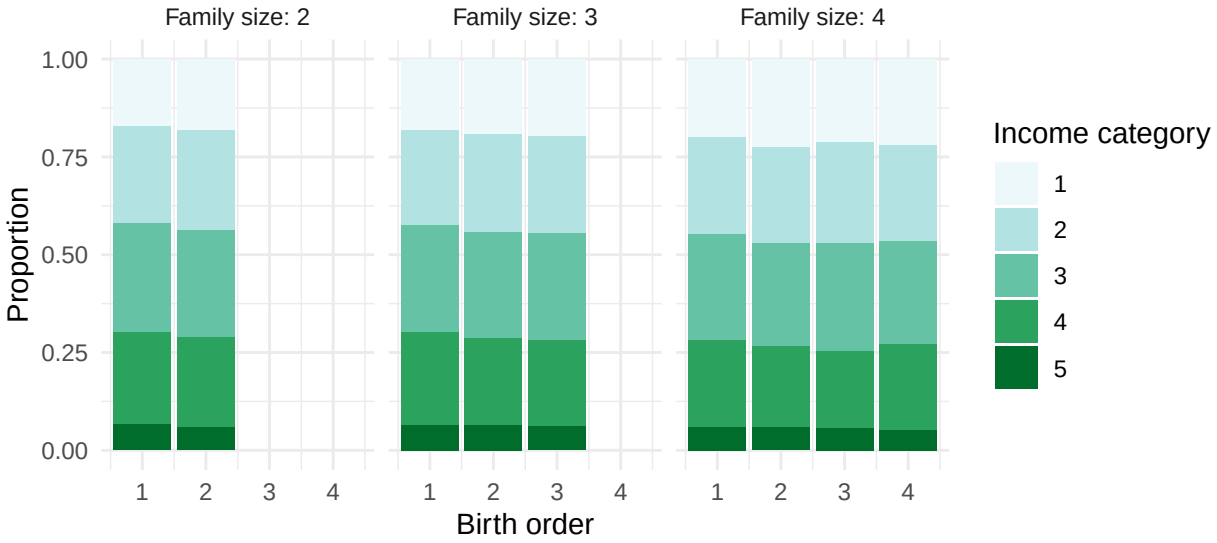


Figure 6: Income by birth order and family size

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 removes birth order and includes university attendance. Column 3 includes birth order and university attendance. The coefficient on university attendance barely shifts, but the coefficient on birth order decreases by about 20%. Columns 4 and 5 repeat this exercise with income category. Again, the coefficient on income is unchanged by including birth order but birth order's effect decreases when income is included.

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.[4] 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 4 in the appendix reruns regressions controlling for these scores. Results are almost unchanged.

5 Conclusion

TODO: Write!

[4]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 Proof of Proposition 1

Write b_t, g_t, s_t, n_t for the population proportion of each type in generation $t = 1, 2, ...$ Sometimes we drop the subscript. Since G_i and S_i are independent in generation 1,

$$\begin{aligned} b_1 &= \gamma\sigma; \\ g_1 &= \gamma(1-\sigma); \\ s_1 &= (1-\gamma)\sigma; \\ n_1 &= (1-\gamma)(1-\sigma). \end{aligned}$$

Within either group $J \in \{H, L\}$, write the proportion of types in generation t as b_J, g_J, s_J, n_J for the respective types.

Dividing children up by their parents' types, we can calculate the proportion with each of the four types:

Summing these in each group, and using that $n_H = 0$ and $b_L = 0$, gives the proportion of children of each type in each group:

	b_J	g_J	s_J	n_J
b_J	$1;0;0;0$	$\frac{1}{2};\frac{1}{2};0;0$	$\frac{1}{2};0;\frac{1}{2};0$	$\frac{1}{4};\frac{1}{4};\frac{1}{4};\frac{1}{4}$
g_J	$\frac{1}{2};\frac{1}{2};0;0$	$0;1;0;0$	$\frac{1}{4};\frac{1}{4};\frac{1}{4};\frac{1}{4}$	$0;\frac{1}{2};0;\frac{1}{2}$
s_J	$\frac{1}{2};0;\frac{1}{2};0$	$\frac{1}{4};\frac{1}{4};\frac{1}{4};\frac{1}{4}$	$0;0;1;0$	$0;0;\frac{1}{2};\frac{1}{2}$
n_J	$\frac{1}{4};\frac{1}{4};\frac{1}{4};\frac{1}{4}$	$0;\frac{1}{2};0;\frac{1}{2}$	$0;0;\frac{1}{2};\frac{1}{2}$	$0;0;0;1$

Table 2: Proportions of children born from different types of parents (b-type; g-type; s-type; n-type)

$$
\begin{aligned}
b_{J,t+1} &= b_J + \frac{1}{2}\rho_J \\
g_{J,t+1} &= g_J - \frac{1}{2}\rho_J \\
s_{J,t+1} &= s_J - \frac{1}{2}\rho_J \\
n_{J,t+1} &= n_J + \frac{1}{2}\rho_J
\end{aligned}
$$

where

$$\rho_J \equiv g_J s_J.$$

In words, when a group contains both g-types and s-types, some of them marry each other and produce some b- and n-type offspring: this increases the correlation between G_i and S_i.

Summing over the groups, we can write the new generation's proportion of types in the whole population as

$$
\begin{aligned}
b_{t+1} &= b_t + \frac{1}{2}\rho \\
g_{t+1} &= g_t - \frac{1}{2}\rho \\
s_{t+1} &= s_t - \frac{1}{2}\rho \\
n_{t+1} &= n_t + \frac{1}{2}\rho
\end{aligned}
$$

where

$$\rho \equiv \theta\rho_H + (1-\theta)\rho_L. \tag{2}$$

We can now calculate the covariance between G and S in the second generation. The covariance is just $E(G_iS_i) - E(G_i)E(S_i)$; in general, at generation t, this is $b_t - \gamma\sigma$. For $t = 2$ this is just equal to ρ.

In the first generation, we can calculate:

$$
\begin{aligned}
b_{H,1} &= \gamma\sigma/\theta; \\
g_{H,1} &= (1-k)\phi/\theta; \\
s_{H,1} &= k\phi/\theta; \\
n_{H,1} &= 0;
\end{aligned}
$$

where

$$\phi \equiv \theta - \gamma\sigma.$$

Similarly:

$$\begin{aligned} b_{L,1} &= 0; \\ g_{L,1} &= \frac{\gamma(1-\sigma) - \phi(1-k)}{1-\theta} \\ s_{L,1} &= \frac{(1-\gamma)\sigma - \phi k}{1-\theta} \\ n_{L,1} &= \frac{(1-\gamma)(1-\sigma)}{1-\theta}; \end{aligned}$$

and from these

$$\begin{aligned} \rho_H &= \frac{\phi^2 k(1-k)}{\theta^2}; \\ \rho_L &= \frac{[\gamma(1-\sigma) - \phi(1-k)][(1-\gamma)\sigma - \phi k]}{(1-\theta)^2}. \end{aligned}$$

Hence:

$$\begin{aligned} \rho = &\frac{\phi^2 k(1-k)}{\theta} + \\ &\frac{[\gamma(1-\sigma) - \phi(1-k)][(1-\gamma)\sigma - \phi k]}{(1-\theta)}. \end{aligned}$$

Note that when $k = 1$ and $\theta = \sigma$, ρ evaluates to 0. Similarly, when $k = 0$ and $\theta = \gamma$, ρ evaluates to 0. This confirms the claims in the main text about the two extreme example societies.

ρ is a quadratic in k. Differentiating ρ by k gives

$$\begin{aligned} \frac{d\rho}{dk} &= \frac{1}{\theta}\phi^2(1-2k) + \frac{1}{(1-\theta)}[\phi^2(1-2k) + \phi(\sigma-\gamma)] \\ &= \frac{1}{\theta(1-\theta)}\phi^2(1-2k) + \frac{1}{(1-\theta)}\phi(\sigma-\gamma) \\ &= \frac{1}{(1-\theta)}\left[\frac{\phi^2}{\theta}(1-2k) + \phi(\sigma-\gamma)\right]. \end{aligned} \tag{3}$$

This is decreasing in k, proving that ρ is concave.

We find the maximum for ρ by solving the first order condition. In (3), setting the term in square brackets to 0 gives:

$$k = \frac{1}{2} + \frac{1}{2}\frac{\theta}{\theta - \gamma\sigma}(\sigma - \gamma). \tag{4}$$

We need to show this is strictly between 0 and 1 as θ approaches σ or γ.

First assume $\sigma < \gamma$. As $\theta \to \sigma$, the expression (4) approaches $k = \frac{1}{2} + \frac{1}{2}\frac{(\sigma-\gamma)}{(1-\gamma)}$. The requirement for this to be interior is that $|\sigma - \gamma| < |1 - \gamma|$. For $\sigma < \gamma$ this is equivalent to $\gamma < (1+\sigma)/2$, which holds by condition (1).

Next assume $\gamma < \sigma$. As $\theta \to \gamma$, (4) approaches $k = \frac{1}{2} + \frac{1}{2}\frac{(\sigma-\gamma)}{(1-\sigma)}$. This is interior when $|\sigma - \gamma| < |1 - \sigma|$, equivalently when $\sigma < (1+\gamma)/2$, which again holds by (1).

6.2 Proof of Proposition 2

We can rewrite table 2 as:

	b_J	g_J	s_J	n_J
b_J	$1;0;0;0$	$\alpha;1-\alpha;0;0$	$\frac{1}{2};0;\frac{1}{2};0$	$\frac{\alpha}{2};\frac{1-\alpha}{2};\frac{\alpha}{2};\frac{1-\alpha}{2}$
g_J	$\alpha;1-\alpha;0;0$	$0;1;0;0$	$\frac{\alpha}{2};\frac{1-\alpha}{2};\frac{\alpha}{2};\frac{1-\alpha}{2}$	$0;\frac{1}{2};0;\frac{1}{2}$
s_J	$\frac{1}{2};0;\frac{1}{2};0$	$\frac{\alpha}{2};\frac{1-\alpha}{2};\frac{\alpha}{2};\frac{1-\alpha}{2}$	$0;0;1;0$	$0;0;\frac{1}{2};\frac{1}{2}$
n_J	$\frac{\alpha}{2};\frac{1-\alpha}{2};\frac{\alpha}{2};\frac{1-\alpha}{2}$	$0;\frac{1}{2};0;\frac{1}{2}$	$0;0;\frac{1}{2};\frac{1}{2}$	$0;0;0;1$

Table 3: Proportions of children born, α-rule for social inheritance

TODO: complete!

6.3 Regressions controlling for polygenic scores

Table 4: 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.