Exploring the correlation between per-SNP effects on fertility and on education among rare and common alleles*

David Hugh-Jones

4/12/23

Genetics for educational attainment have been selected against in modern populations. Observed effect sizes are small, but are substantively large after correcting for errors in variables. This correction depends on whether, of genetic variation driving educational attainment, the as-yet-unmeasured part has the same relationship to fertility as the measured part. To check whether the education/fertility relationship is the same among rare and common alleles, I use per-allele summary statistics and implement a correction for errors in variables. The education/fertility correlation is about 50% smaller among rare alleles. However, simulations show that the errors-in-variables correction gets less accurate for minimum allele frequencies below 0.05. We cannot yet be sure of the effect size of natural selection on contemporary humans, so more research is needed.

Early draft. Don't quote me on anything.

Many polygenic scores are undergoing natural selection in contemporary advanced societies. In particular, lower scores for educational attainment (EA) are being selected for. However, effect sizes are small. Hugh-Jones and Abdellaoui (2022) estimate that the PGS for EA from Lee et al. (2018) (EA3) was reduced by 0.03 standard deviations in children of the UK Biobank generation, with a similar or slightly smaller reduction in their parents' generation. These are not large effects: a reduction of 0.03 standard deviations means that 51.2% of the child generation were below the mean of the parent generation.

*Currently jobless. Email davidhughjones@gmail.com

PGS are created from summary statistics which are estimated with noise. This means that when we regress fertility on a PGS, its effects will be smaller than the effects of the "true polygenic score". This is an errors-in-variables problem. The heritability of EA is about 40%, but in the UK Biobank sample, EA3 only explains about 4.5% of variance in EA. Under standard errors-in-variables assumptions, to find the coefficient of "true PSEA", the true best predictor of EA from genetic data, we should multiply our estimate by the ratio of these variances, giving 0.03 x 40/4.5 = 0.26. A reduction of 0.26 standard deviations in the true measure of EA in one generation would mean that 60.3% of the child generation were below the mean of the parent generation. This would be socially and economically significant.

This calculation assumes that the relationship between effects on EA and fertility is the same in the unmeasured part of true PSEA as in the measured part. That might not be the case. For example, EA3 and subsequent polygenic scores such as EA4 are estimated using common SNPs which are captured on DNA array chips. Unmeasured PSEA is likely to be partly a result of rare variants or *de novo* mutations. While common alleles which lower EA raise fertility on average, rare alleles and new mutations might simultaneously harm EA and fertility. Or, the relationship may simply be weaker among rare alleles.

While by definition we can't yet learn about the effects of unmeasured PSEA, we can look at existing SNPs to see if the correlation between effects on EA and on fertility is the same among more common and rarer alleles. If it is the same, then that will increase our confidence that the same relationship holds for as-yet-undiscovered variants. If it is not the same, then we will not be so sure.

I downloaded summary statistics for EA4 (Okbay et al. 2022) and for number of children born (NCB) (Barban et al. 2016) from https://thessgac.com. I merged the datasets, discarding SNPs that were not in both. I split the SNPs into groups, by the reported standard error of the EA4 effect size estimate. Within each group, I regressed alleles' NCB betas on their EA4 betas. Figure 1 plots regression coefficients for each group against the mean MAF within each group. The EA-NCB relationship is apparently weaker for rarer alleles.

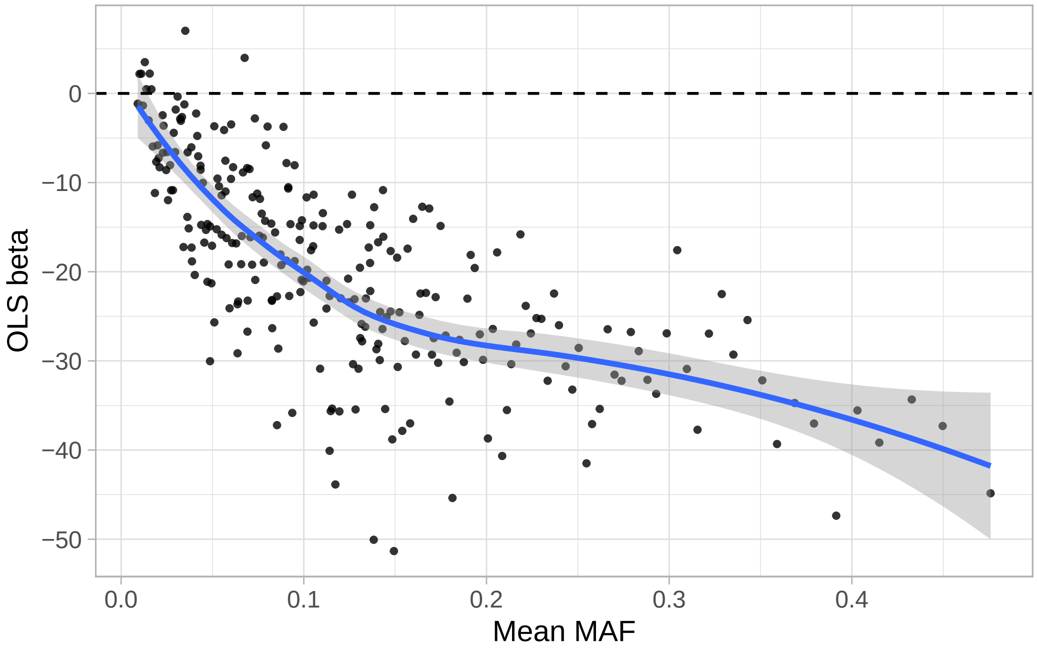


Figure 1: Regressions of NCB betas on EA4 betas, by mean MAF within each group. The blue line is a loess smoother.

However, this result could happen mechanically. The per-SNP summary statistics are estimated with error, and this error is larger for rarer alleles, since they are estimated with fewer cases. As a result, we again have an errors-in-variables problem. An unbiased estimate of the true relationship between EA4 effects and fertility will be

$$\beta = \hat{\beta}\frac{\sigma_X^2 + \sigma_\eta^2}{\sigma_X^2}$$

where $\hat{\beta}$ is the OLS estimate, σ_X^2 is the variance of the distribution of true effect sizes and σ_η^2 is the variance of the error term in the estimated effect sizes.

Within each group, I estimated $\sigma_X^2 + \sigma_\eta^2$ by the variance of the group's estimated EA4 effect sizes, and σ_η^2 from the mean of the squared standard error of the EA4 effect size estimates (as reported, after adjustment for stratification).[1] I then calculated the corrected β for each group using the formula above.

[1] If σ_η^2 was larger than my estimate of $\sigma_X^2 + \sigma_\eta^2$, then I discarded the group as containing too little information to be useful.

```
Warning: Removed 3 rows containing non-finite values (`stat_smooth()`).
```

```
Warning: Removed 3 rows containing missing values (`geom_point()`).
```

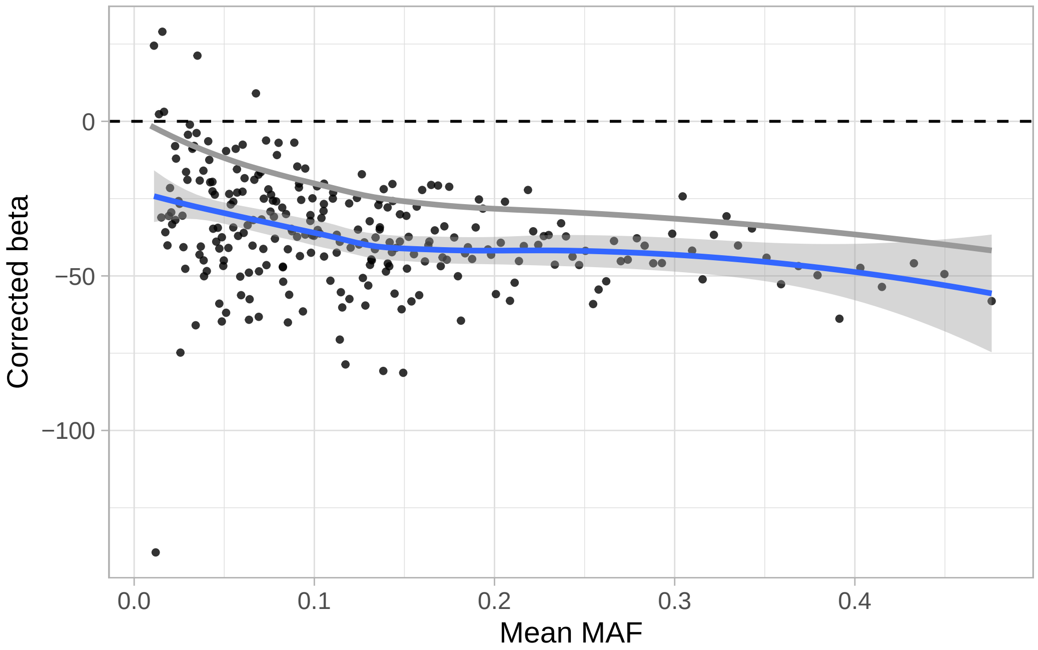


Figure 2: Regressions of NCB betas on EA4 betas, by mean MAF within each group. Coefficients corrected for errors in variables. The blue line is a loess smoother. The grey line shows the uncorrected smoother from the previous figure.

Figure 2 plots the corrected β. The correction indeed makes effects of rare alleles absolutely bigger. Nevertheless, there remains a clear negative relationship between MAF and effect size. In a regression of effect size on MAF we can reject the null at $p = 1.065 \times 10^{-6}$. The intercept is -28.595 and the slope is -58.464, suggesting that rare alleles (MAF ≈ 0) will have about half the EA/fertility correlation of common alleles (MAF $= 0.5$).

However, we have to interpret this cautiously. Simulations (in the appendix) show that the errors-in-variables correction becomes less accurate for small values of MAF.[2] Also, the lowest MAF in the data is 0.001 – which implies millions of carriers worldwide. There are obvious risks in extrapolating to rarer alleles.

[2] But note that if we exclude groups with a mean MAF below 0.1, there is still a significant slope of mean MAF on β of -26.626 ($p = 0.039$).

These results mostly suggest that more research is needed. The EA/fertility relationship is smaller for rarer SNPs, but not vastly so, and we cannot be sure what happens at very rare alleles. Ultimately the best way to gauge the size of natural selection effects will be to create more accurate polygenic scores for EA and other phenotypes, and relate them directly to fertility.

Appendix: simulations

The correction method makes the following approximations:

- The variance of estimated EA4 effect sizes is estimated by pooling estimates within each group.
- The variance of errors in effect sizes is estimated by the squared mean of the reported standard errors within each group.

This may be inaccurate if the distributions vary within a group.

To check that the correction method worked, I ran simulations:

- I drew a simulated true effect on EA4 for each SNP in the data, and added normal noise with the reported standard error for the EA4 effect size to create a simulated observed effect.
- I created a simulated true effect on NCB, related to the true effect on EA4 and/or to the MAF; and added normal noise to create a simulated observed effect.
- I split SNPs by their EA4 effect size standard error. I used 1000 quantile groups where quantile $q_n = (n/1000)^2$, i.e. smaller groups for smaller standard errors.
- I regressed observed NCB effects on observed EA4 effects within groups.
- I corrected the regression betas as in the main text.

Figure 3 shows the results for different specifications of the relationship between MAF and the EA4/NCB coefficient: constant, linear and nonlinear. The quantile groups were chosen by trial and error, to balance out 2 competing sources of error. A smaller N in each group increases the error of the uncorrected beta estimate, and the error of the estimate of σ^2_X. On the other hand, a smaller N has a smaller range of reported standard errors of the EA4 betas, which makes the approximation of σ^2_η by its group mean more accurate. Simulations by the chosen method are roughly accurate down to about MAF = 0.05, but become noisier below that point.

```
Warning: Removed 1 rows containing non-finite values (`stat_smooth()`).
```

```
Warning: Removed 1 rows containing missing values (`geom_point()`).
```

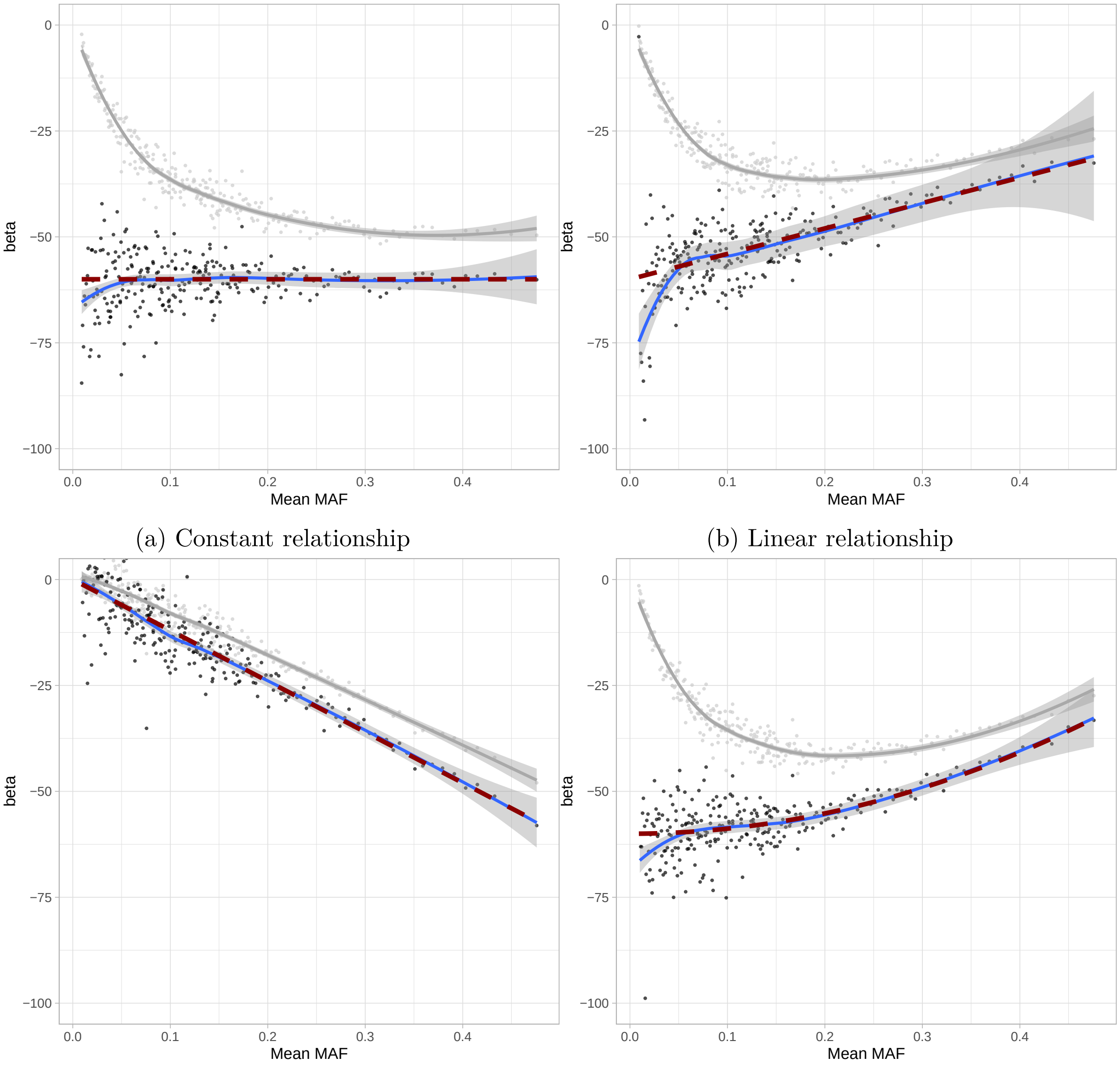
References

Barban, Nicola, Rick Jansen, Ronald De Vlaming, Ahmad Vaez, Jornt J Mandemakers, Felix C Tropf, Xia Shen, et al. 2016. "Genome-Wide Analysis Identifies 12 Loci Influencing Human Reproductive Behavior." *Nature Genetics* 48 (12): 1462–72.

Hugh-Jones, David, and Abdel Abdellaoui. 2022. "Human Capital Mediates Natural Selection in Contemporary Humans." *Behavior Genetics* 52 (4-5): 205–34.

Lee, James J, Robbee Wedow, Aysu Okbay, Edward Kong, Omeed Maghzian, Meghan Zacher, Tuan Anh Nguyen-Viet, et al. 2018. "Gene Discovery and Polygenic Prediction from a Genome-Wide Association Study of Educational Attainment in 1.1 Million Individuals." *Nature Genetics* 50 (8): 1112–21.

Okbay, Aysu, Yeda Wu, Nancy Wang, Hariharan Jayashankar, Michael Bennett, Seyed Moeen Nehzati, Julia Sidorenko, et al. 2022. "Polygenic Prediction of Educational Attainment Within and Between Families from Genome-Wide Association Analyses in 3 Million Individuals." *Nature Genetics* 54 (4): 437–49.



(a) Constant relationship

(b) Linear relationship

(c) Linear relationship

(d) Nonlinear relationship

Figure 3: Simulations. Regressions of simulated observed NCB betas on simulated observed EA4 betas. Each dot represents a regression using SNPs having EA4 effect size standard errors within a given interval. Grey dots show the uncorrected betas. Black dots are the corrected betas. The red dashed line shows the simulated true relationship. The blue line is a loess smoother. The grey line is a smoother for the uncorrected betas, for comparison. 95% confidence intervals are shown.