# Natural Selection Across Three Generations of Americans

David Hugh-Jones Tobias Edwards

2024-02-01

We investigate natural selection on polygenic scores in the contemporary US, using the Health and Retirement Study. Results partially support the economic theory of fertility as an explanation for natural selection: scores which correlate negatively (positively) with education are selected for (against). The theory predicts that selection coefficients should be stronger among low-income, less educated, unmarried and younger parents, but these predictions are only partly borne out: coefficients are larger among low-income and unmarried parents, but not among younger parents or those with less education. Using respondents' number of siblings and grandchildren, we examine selection across three generations, finding consistent results over time. We also estimate effect sizes corrected for noise in the polygenic scores. Selection for some health traits is similar in magnitude to that for cognitive traits.

Hugh-Jones and Abdellaoui (2022) explain patterns of natural selection on polygenic scores in the UK, using an economic theory of fertility derived from G. S. Becker and Tomes (1976). The theory has two components:

1. There is a trade-off between time spent working and raising children. This "substitution effect" leads people with more human capital and higher expected wages to have fewer children. Evidence for this is that polygenic scores which correlate positively with human capital correlate negatively with number of children, i.e. they are being selected against; conversely, scores which correlate positively with human capital are being selected for.
2. The trade-off is sharper for low-income people, people with low human capital, and single parents. Because these groups value income more at the margin, the substitution effect is stronger for them. In other groups, the substitution effect is balanced by the "income effect", that children become more affordable when you get richer. As a result, natural selection is stronger among these groups. Evidence for this is that

scores' regression coefficients on number of children are larger among people with lower income or less education, and single parents.

Here, we make an independent test of the theory in the US population, using the Health and Retirement Survey (HRS 2023a, 2023b). The HRS is more representative of the population than UK Biobank, which addresses one potential weakness of the previous paper. Using information on respondents' siblings and grandchildren, we can also extend the analysis to three generations of Americans. To preview our results, we confirm point 1 above, but only see partial and ambiguous support for point 2. (TODO: or can we be stronger?)

## Data

The HRS sample focuses on cohorts born between 1920 and 1960, but contains some younger and older participants. we include only male participants born before 1965 and female participants born before 1970, which guarantees that most will have completed their fertility by 2010. The resulting sample contains 8827 genotyped white participants. We focus on these because the sample size is large enough. The appendix reports the most basic analyses for the 2319 genotyped black participants.

Genotyping took place in 2006, 2008 and subsequent years. PGS were taken from those pre-calculated by the HRS (Ware et al. 2020) and those produced by the Social Science Genetic Association Consortium, as part of their Polygenic Index Repository (J. Becker et al. 2021). Scores created by the HRS were provided for black and white participants, but Polygenic Index Repository scores were only created for white participants.

For the white participants, when scores from the two samples measured the same trait, we only used the PGS from the Polygenic Index Repository. For some traits, polygenic scores were created from both European ancestry and multiple ancestry GWAS. We choose to use polygenic scores trained only on individuals of European ancestry. We discard obsolete PGS for which there is a newer, more accurate score targeting the same phenotype. we also discard PGS for number of children ever born (but keep scores for age at first birth). This leaves a total of 68 scores for the white participants. PGS are rescaled to zero mean and unit variance. In all regressions using PGS, we control for ten within-ethnicity principal components of the DNA array data. For our purposes, this is equivalent to residualizing on principal components (Wurm and Fisicaro 2014). Effectively, our analysis is of the residualized PGS that are widely used in practical applications. [TODO actually we could just residualize; only the SE of the PCs is affected]

The key dependent variable is relative lifetime reproductive success (RLRS): number of children ever born, divided by the mean number of children of people born in the same year. RLRS is calculated pooling ethnicities, i.e. treating them as members of the same

biological population. The mean number of children of people born in the same year was calculated using sampling weights.

The HRS contains weights which match survey respondents to the US population. We use weights for the biomarker subsample (*BIOWGTR in the HRS tracker file). Since half the sample enters the extended interview including biomarker data in each biannual survey, we weight individuals by either their 2010 weight or their 2012 weight. This maximizes the available sample of both black and white respondents, and should approximately match the US population of the sample cohorts between 2010 and 2012. Statistical tests are adjusted for clustering and stratification using the R "survey" package (Lumley 2023).

## Results

We estimate coefficients of PGS on RLRS. These are not meant to identify causal effects; recall that natural selection involves correlation, not necessarily causation, between selected characteristics and fertility. Appendix Figure 6 shows coefficients. Standard errors are large because of the relatively low sample sizes. 3 scores are significant at Bonferroni-corrected $p < 0.05/68$. The scores are age at first birth, educational attainment, ADHD, and self-rated health. But we are most concerned with looking at patterns across scores rather than judging the significance of individual scores.

Figure 1 plots each PGS's regression beta on RLRS against its regression beta on educational attainment. The relationship is negative (correlation -0.819, bootstrap 95% C.I. -0.99 to -0.648). Survey bootstraps (Canty and Davison 1999) are used so as to make inferences from the sample of respondents.

We can also examine natural selection in the preceding and succeeding generations, by using reported number of siblings and grandchildren respectively. We regress PGS on respondents' number of living siblings in 2010. Data for dead siblings has too many missing values to use. We reweight respondents by the reciprocal of their number of siblings, to account for parents of many siblings being more likely to be a parent of a respondent. Parents of no siblings cannot be included, so coefficient sizes are not comparable across the generations. Appendix Figure 7 plots coefficients on number of siblings versus coefficients on years of education. Correlations are significant and negative (correlation -0.387, bootstrapped 95% C.I. -0.691 to -0.084). There is a positive and significant correlation across generations, i.e. between PGS coefficients on number of siblings and number of children (0.557, bootstrapped 95% C.I. 0.329 to 0.785).

To examine selection in the next generation, we divide the respondents' number of grandchildren by their number of children. In other words, we calculate the average number

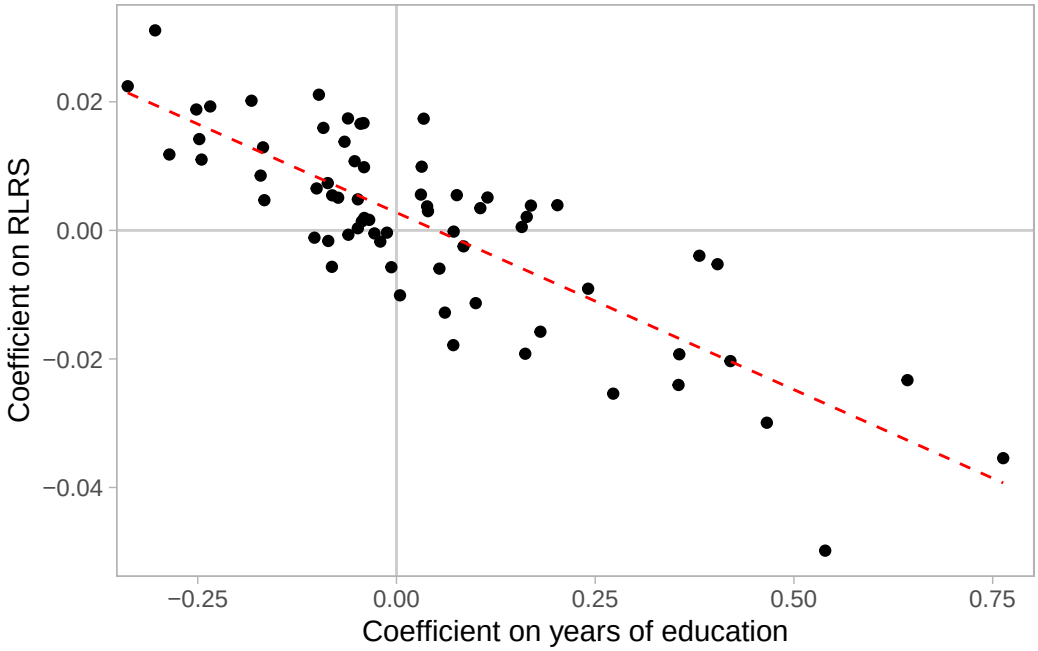


Figure 1: Scatterplot of PGS regression coefficients on RLRS against coefficients on years of education. Each dot is one polygenic score. Controls include 10 principal components of genetic array data. Dashed lines are fitted from linear regressions.

of children of the respondent's children. This indicates reproductive success in the second generation, which we call RLRS2. We removed 216 respondents who report having grandchildren despite reporting having reported no children. In regressions we reweight respondents by the number of children they have, since more fecund grandparents account for a larger proportion of the next generation.

Older grandparents will have more time for their number of grandkids to accumulate. To deal with this time trend, we subset the data to parents whose oldest child was 40 years old or older in 2016, when the number of grandchildren was recorded. Appendix Figure 5 shows the distribution of the birth year of the oldest child, adjusted for sampling weights. A majority of the observations are after 1960, the most recent year the respondents were born in. This reassures us that we are observing reproduction in later generations.

We regress the respondent's (the grandparent's) PGS on RLRS2. The resulting coefficient is only a proxy for selection in the next generation, not a perfect measure. The grandparent's PGS indicates, but does not determine the PGS of the parent. To know the expected parent's PGS, we would need both grandparents' PGS. Given that some respondents do not have a partner in the HRS and some have had children with multiple partners, that approach is untenable in our sample. We expect the effect of the grandparent's PGS to also depend upon the level of assortative mating for the trait. For traits with high assortative mating, the grandparent's PGS will more strongly predict the parent's PGS, leading to a greater regression slope.

Appendix Figure 8 plots coefficients on RLRS2 versus coefficients on years of education. Correlations are significant. Standard errors are large (correlation -0.383, bootstrapped 95% C.I. -0.722 to -0.044). There is a positive correlation between PGS coefficients on RLRS2 and RLRS (0.444, bootstrapped 95% C.I. 0.099 to 0.789), indicating some stability in selection over time.

## Differences across social groups

We next test part 2 of the theory by interacting PGS with measures of education, income, marital status, and age at first birth. Education is years of education, split at 12 years. Income is respondent's mean wage income over all surveys, residualized on a full set of birth year dummies, and median-split. We call the low-education, low-income, unmarried or younger-AFB group the "disadvantaged group".

In Hugh-Jones and Abdellaoui (2022) we could simply look for significant differences between coefficients on RLRS among the groups, for each polygenic score. That will not work here because of the smaller sample size. At the same time, scores are not independent and not a sample from a population. To get round this, we test the null hypothesis that there

are no differences in coefficients between social groups for any PGS, against the alternative that for some PGS, coefficients are larger for the disadvantaged group. Here, "larger" means in the direction predicted by the score's correlation with education: more positive for scores which are negatively correlated with education, and more negative for scores which are positively correlated with education. So, we run regressions of the form

$$RLRS_i = \alpha + \beta PGS_i + \gamma DIS_i + \delta(PGS_i \times DIS_i) + \varepsilon_i$$

where $DIS_i$ is a dummy for $i$'s membership in the disadvantaged group and where $PGS_i$ has been sign-flipped to correlate negatively with education. Under the null, the estimated $\delta$s will be distributed around zero; in expectation half will be positive and half negative. Under the alternative, more than half the $\delta$s will be positive. We bootstrap respondents and count the number of positive $\delta$s in each resample. If 95% confidence intervals are above 68/2 = 34, we reject the null in favour of the alternative. [TODO: what if some are negative and some are positive? I think this is fine because it makes our test conservative? Maybe??]

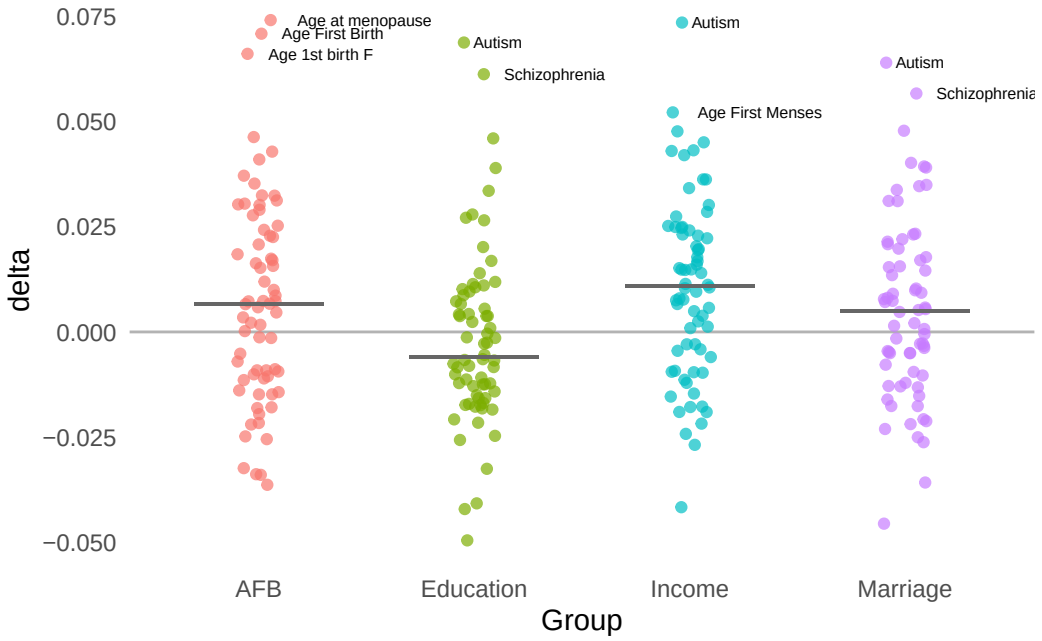


Figure 2: Coefficients of interaction terms ($\delta$) for effects of polygenic scores on RLRS among different social groups. A positive $\delta$ means that the coefficient was larger for the disadvantaged group, where the score was sign-flipped to correlate negatively with education. Horizontal lines show medians.

Figure Figure 2 plots values of $\delta$ for each interaction term, without bootstrapping. The majority of $\delta$s are positive for age at first birth, income and marriage but not for education. In other words, for younger parents, poorer participants, and unmarried participants, most PGS have larger effects on RLRS, but this is not true for less educated participants.

Table 1 shows estimates and bootstrapped 95% confidence intervals for the number of scores where the $\delta$ term is positive. Confidence intervals exclude 34 only for income.

Table 1: Estimates and bootstrap 95% confidence intervals for numbers of PGS out of 68 where the disadvantaged group had a larger selection coefficient than the advantaged group. Groups are: 0-12 years education vs. 13-17 years; below vs. above median income; all others vs. married; below vs. above median age at first birth; born before vs. after 1942. 199 bootstraps.

| | Estimate (95% C.I.) |
|---|---|
| Education | 28 (17.3 to 38.7) |
| Income | 47 (39.6 to 54.4) |
| Marriage | 38 (29.1 to 46.9) |
| Age at first birth | 41 (31.6 to 50.4) |
| Birth year | 28 (18.7 to 37.3) |

Why does the US data show fewer differences by socio-economic status (SES) than the UK? One possibility is that SES maps on to race in the US, so that ethnic differences here capture some of the variation seen in the UK. The regression coefficient of phenotypic educational attainment on RLRS is more negative among black than white respondents (blacks: -0.076, 95% C.I. -0.093 to -0.058; whites: -0.029, 95% C.I. -0.035 to -0.023; cf. Goldscheider and Uhlenberg (1969), Johnson (1979), Yang and Morgan (2003)). And the slope of PGS education coefficients on fertility coefficients is larger among black respondents, though imprecisely estimated. But comparisons of PGS selection coefficients between the ethnic groups are hard because of the smaller sample size and differences in the scores' predictive power, so this hypothesis can only be speculative. Looking within the sample, changes in PGS by birth year are small for both groups, and are probably mostly driven by selective mortality. Another possibility is that the US cohort were exposed to a smaller welfare state than the UK cohort, since many of them had children before the "Great Society" programs of the 1960s. The last line of Table 1 shows that effect sizes are larger for respondents born after 1942, but the difference is imprecisely estimated and not significant.

Lastly, we would like to know natural selection's effect sizes. The bivariate correlation of PGS with RLRS gives the change in one generation in the PGS due to natural selection, measured in standard deviations. Polygenic scores contain error, so estimated correlations

are biased towards zero compared to the correlation of the true PGS. They can be scaled up by

$$\hat{\beta}_{TRUE} = \hat{\beta}_{PGS}\sqrt{\frac{h^2}{R^2_{PGS}}}$$

where $h^2$ is the heritability of the PGS target phenotype and $R^2_{PGS}$ is the coefficient of determination of the measured PGS on the target phenotype (J. Becker et al. 2021). Moving from error correction in a univariate model to a multivariate model, controlling for principal components of genetic array data, requires a slightly more complex correction, although it uses the same parameters. To do this we use the error-corrected estimator developed by J. Becker et al. (2021).

SNP- or chip-heritabilities and $R^2_{PGS}$ are calculated by J. Becker et al. (2021) for scores in the Polygenic Index Repository. The authors use GCTA to estimate heritability. When the corresponding phenotype is not available in the HRS to estimate $R^2_{PGS}$, we use parameters estimated by the authors in the Wisconsin Longitudinal Study instead. We also perform error correction with twin heritabilites. To attain precise estimates we use heritabilities from a meta-analysis including over 14,000 twin pairs authored by Polderman et al. (2015)[1]. We remove polygenic scores with $R^2_{PGS} < 0.005$ to focus on PGS with adequate power. This removed the ADHD PGS which significantly predicted RLRS, even after bonferroni correction[2].

Figure 3 reports the error corrected estimates of selection. As a rule of thumb, a 0.1 standard deviation change in a polygenic score over a generation might count as "serious": about 54% of the new generation will be below the parents' mean. Many upper confidence bounds meet that threshold, but lower bounds are often small or include zero. The confidence bounds capture uncertainty from sampling variation, but not other sources, including uncertainty about the true $h^2$, the true $R^2_{PGS}$, limitations of the within-sample phenotypes, noise from correlated environments, and biases in the polygenic scores such as assortative mating, population stratification, and gene-environment correlations. For twin-heritability, different relationships with fertility among variants not measured on the chip. Given all this, the

[1]For age of first birth, we could not find an appropriately close trait in Polderman et al. (2015). For this trait we used a twin heritability of 0.15 as estimated in the Midlife in the United States (MIDUS) sample (Briley, Tropf, and Mills 2017).

[2]J. Becker et al. (2021) estimate $R^2_{PGS}$ for ADHD using binary, yes or no, items of ADHD symptoms. These were PV001-PV018 in the 2016 core survey. They estimated an $R^2_{PGS}$ of 0.003. When we attempt to replicate it we find an $r_{PGS}$ of 0.079, implying an $R^2_{PGS}$ of 0.006. Although this would pass our requirement of $R^2_{PGS} > 0.005$, the correlation was not significant (p = 0.051) and the confidence intervals were large 95% C.I. $-1.719 \times 10^{-4}$ to 0.156 owing to the fact that few respondents were given the ADHD items N = 621. Given the great uncertainty over the true correlation we are unable to justify using the error-corrected estimator with the ADHD polygenic score.

estimates mostly show the limits of our knowledge, and should be treated as best guesses only.

Results from a few traits show substantial uncertainty regarding the true effect size. For asthma, when chip heritability is used, the effect is positive and significant, but negative and significant when twin heritability is used! Such a large disparity is because J. Becker et al. (2021) estimate the chip heritability at 0.015 and we use a twin heritability 0f 0.55. For personality traits, chip heritability is often low (e.g. $< 0.05$) but twin heritabilities are often much higher ($\approx 0.4$) implying substantial missing heritability. For these traits, the confidence intervals on twin heritabilities are enormous. Error-corrected effect sizes and the $R^2_{PGS}$ and $h^2$ parameters used can be found on the Github page for this paper.

To estimate how natural selection will contribute to changing the phenotype, we can multiply the change in the mean genetic value by the correlation between genetic values and phenotypes $h$, which is the square root of heritability. For cognitive performance we have estimated a genetic change of $-0.066$ standard deviations per generation, assuming a twin heritability of 0.51. This implies a phenotypic change of $-0.066 \times 0.51^{\frac{1}{2}} = -0.047$ standard deviations per generation, equivalent to $-0.71$ points in the units of IQ, where a standard deviation is equal to 15 points. This calculation assumes the heritability of the trait remains constant, that the genetic correlation across time is equal to one and it ignores the environmental contributions to changes in the phenotype. Given the assumptions required for this calculation, on top of the problems involved in estimating the genetic change, it can only be considered a guess.

## Discussion

The economic theory of fertility is driven by the trade-off between children and income. On the one hand, a higher hourly wage increases the opportunity cost of raising children (a "substitution effect"). On the other hand, higher expected income makes children more affordable (an "income effect"). Prediction 1 of the theory holds when substitution effects dominate income effects overall. Prediction 2 is driven by the specific form of individuals' preferences for income: when utility for income is sharply curved, i.e. marginal utility is high at low incomes but low at high incomes, the substitution effect is stronger for those who expect to earn less.

The results here support prediction 1 but are more ambiguous for prediction 2. PGS coefficients on RLRS also appear larger for low-income groups and unmarried respondents. But there is little evidence for larger coefficients among people with lower education, or younger parents. This may be due to the low sample size. But in the UK, the between-group differences were large (Hugh-Jones and Abdellaoui 2022); differences that big would surely have been visible here. The theory can accommodate this non-result, if preferences

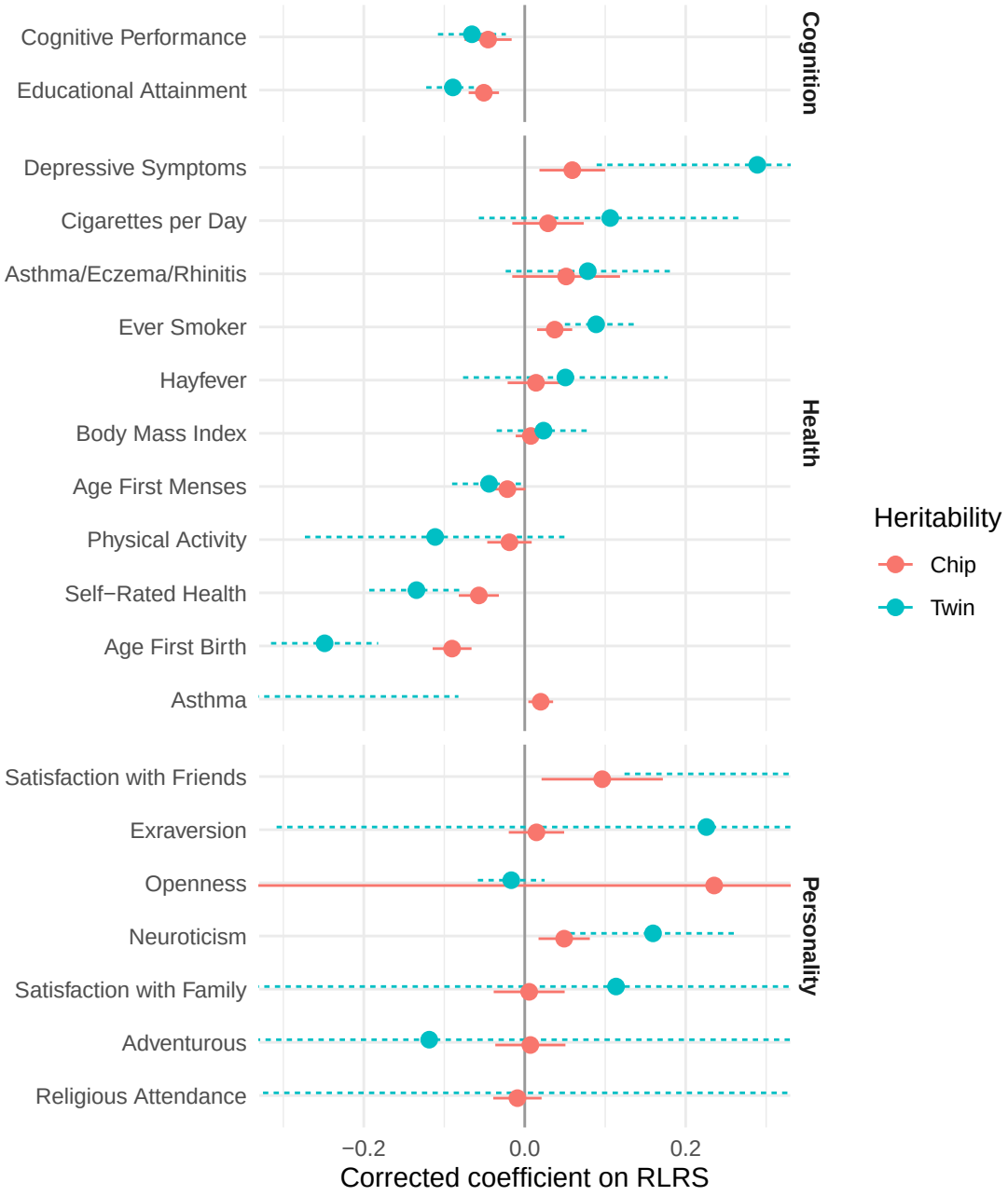


Figure 3: Estimated coefficients of true polygenic scores on RLRS calculated using estimates of chip- and twin-heritability and the most recent polygenic score.

for income are less curved in the US for whatever reason. But note that *any* theory with a negative relationship between education and fertility will give prediction 1.[3] In this sense, results here are less supportive of the economic theory specifically.

We also examined natural selection in different groups. Across three generations of Americans, we found estimates of selection to correlate over time and that selection is consistently more positive for traits that negatively predict educational attainment. Patterns of selection appear consistent across the twentieth and early twenty-first century. The effects of natural selection are accumulating and the underlying causes of selection may have stayed the same over time. The smaller black sample makes most tests inconclusive for this population: we can only say that the data do not reject a negative association between PGS correlations with RLRS and PGS correlations with education.

Lastly, we provide the first estimates of selection differentials of "true" polygenic scores for traits other than intelligence and education, finding results typically between 0-0.2 standard deviations. Our estimate of how much intelligence is changing owing to natural selection, is similar to estimates given in the prior literature. For example, Kong et al. (2017) provide an estimate of $-0.9$ IQ points every 30 years in Iceland, compared to our $-0.71$ points per generation. However, cognitive traits are only one part of the story of natural selection in humans. Whilst past research has focused on these traits, we find selection differentials are often of similar magnitudes for health related traits. The selection differential for self-rated health is greater than those of education and cognitive performance. The most significant, positively selected trait was ADHD, which was also found in the UK (Hugh-Jones and Abdellaoui 2022). Future research should study the health and medical implications of natural selection, in addition to their social implications. To know more, we must await more accurate polygenic scores.

# Appendix


## Acknowledgements

The HRS (Health and Retirement Study) is sponsored by the National Institute on Aging (grant number NIA U01AG009740) and is conducted by the University of Michigan.


Code to reproduce this paper is available at https://github.com/hughjonesd/hrs-selection.

[3] See Balbo, Billari, and Mills (2013) for a broad review of fertility theories.

## Black respondents

There are 2319 genotyped black survey participants. Among this sample, the relationship between polygenic scores' coefficients on RLRS and on education is negative but insignificant and imprecisely estimated (correlation -0.345, bootstrap 95% C.I. -0.734 to 0.044). Correlations between number of siblings and education are insignificant (correlation 0.19, bootstrapped 95% C.I. -0.287 to 0.667). Similarly, coefficients on RLRS2 (children's average number of children) were insignificantly correlated with coefficients on years of education, with large standard errors (correlation -0.007, bootstrapped 95% C.I. -0.545 to 0.531; removed 46 respondents reporting grandchildren but no children).

## Figures

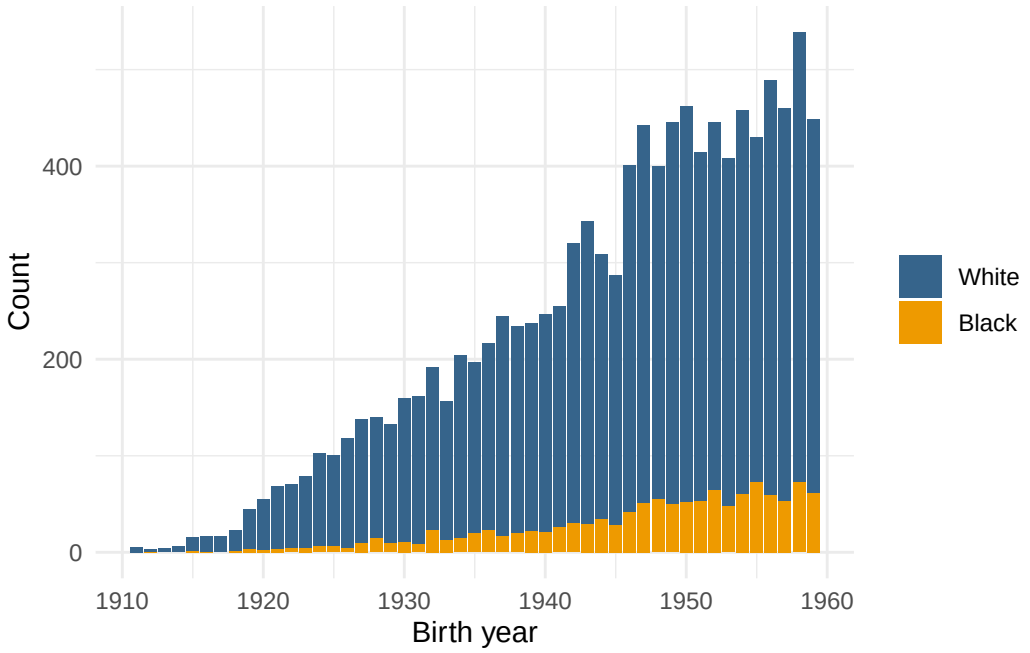


Figure 4: Distribution of birth years adjusted for sampling weights.

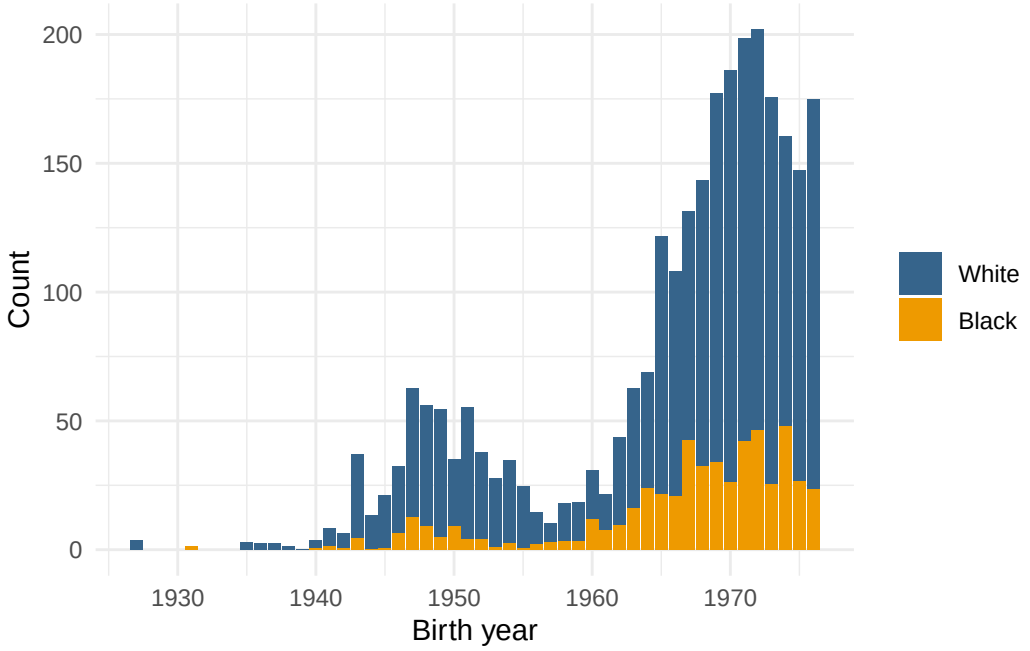


Figure 5: Distribution of birth years for the oldest child of respondents, adjusted for sampling weights.

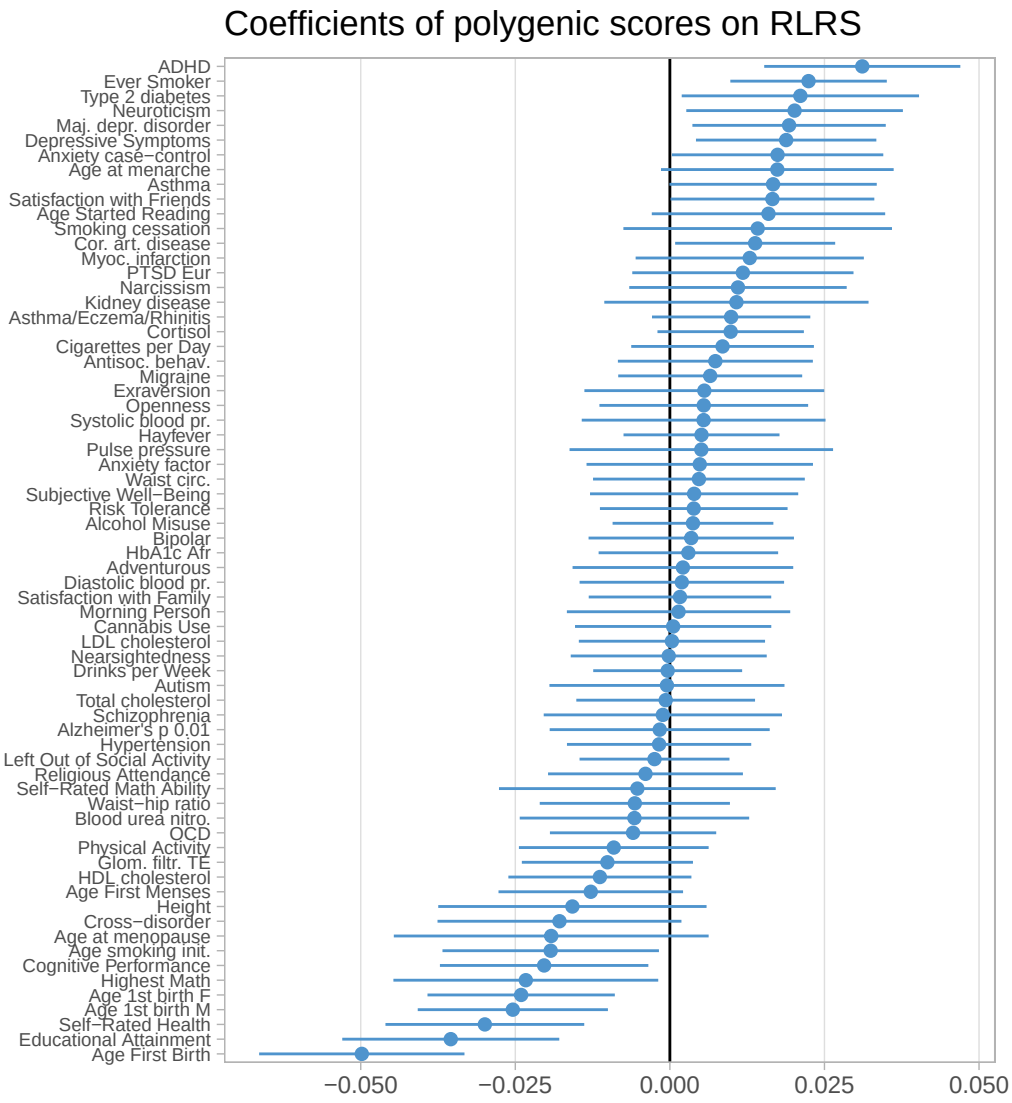


Figure 6: Coefficients of PGS on RLRS, controlling for 10 principal components of genomic array data. Lines are 95% confidence intervals.

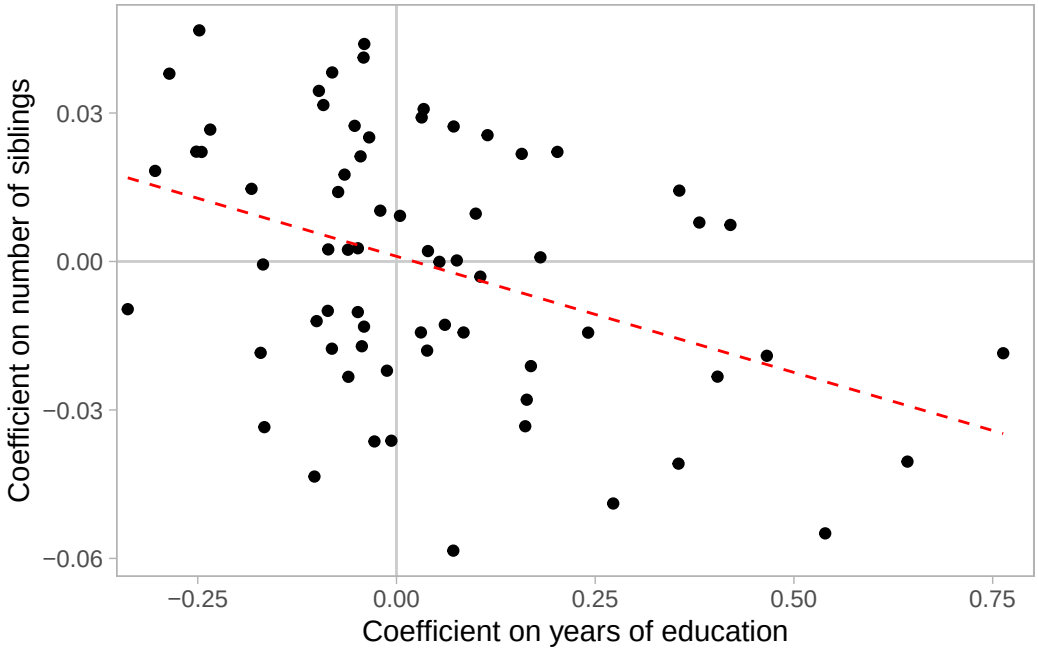


Figure 7: Scatterplot of PGS coefficients on number of live siblings and years of education. Controls include 10 principal components of genetic array data. Dashed lines show linear regressions.

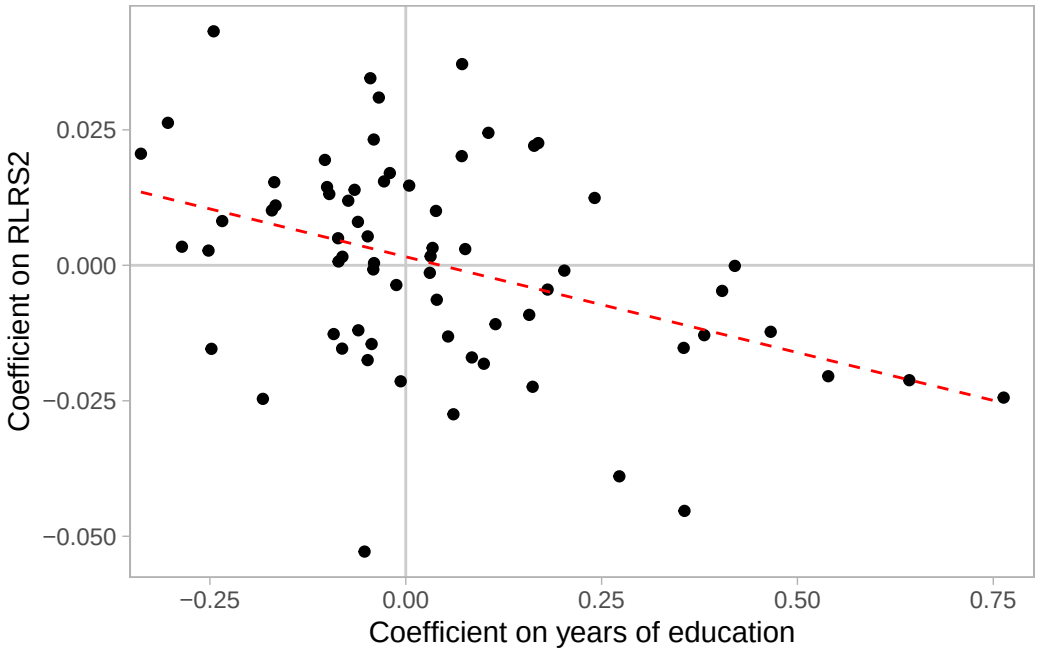


Figure 8: Scatterplot of PGS coefficients on RLRS2 and years of education. RLRS2 is the reproductive success of respondent's offspring relative to the success of other offspring. Controls include 10 principal components of genetic array data. Dashed lines show linear regressions.

## Bibliography


Balbo, Nicoletta, Francesco C Billari, and Melinda Mills. 2013. "Fertility in Advanced Societies: A Review of Research: La fécondité Dans Les Sociétés Avancées: Un Examen Des Recherches." *European Journal of Population/Revue Europeenne de Demographie* 29: 1–38.

Becker, Gary S, and Nigel Tomes. 1976. "Child Endowments and the Quantity and Quality of Children." *Journal of Political Economy* 84 (4, Part 2): S143–62.

Becker, Joel, Casper AP Burik, Grant Goldman, Nancy Wang, Hariharan Jayashankar, Michael Bennett, Daniel W Belsky, et al. 2021. "Resource Profile and User Guide of the Polygenic Index Repository." *Nature Human Behaviour* 5 (12): 1744–58.

Briley, Daniel A., Felix C. Tropf, and Melinda C. Mills. 2017. "What Explains the Heritability of Completed Fertility? Evidence from Two Large Twin Studies." *Behavior Genetics* 47: 36–51.

Canty, Angelo J, and Anthony C Davison. 1999. "Resampling-Based Variance Estimation for Labour Force Surveys." *Journal of the Royal Statistical Society Series D: The Statistician* 48 (3): 379–91.

Goldscheider, Calvin, and Peter R Uhlenberg. 1969. "Minority Group Status and Fertility." *American Journal of Sociology* 74 (4): 361–72.

HRS. 2023a. "Health and Retirement Study - Rand HRS Family Data 2018." Ann Arbor, Michigan: University of Michigan; Survey Research Center, Institute for Social Research.

———. 2023b. "Health and Retirement Study - Rand HRS Longitudinal File." Ann Arbor, Michigan: University of Michigan.

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

Johnson, Nan E. 1979. "Minority-Group Status and the Fertility of Black Americans, 1970: A New Look." *American Journal of Sociology* 84 (6): 1386–1400.

Kong, Augustine, Michael L. Frigge, Gudmar Thorleifsson, Hreinn Stefansson, Alexander I. Young, Florian Zink, Gudrun A. Jonsdottir, et al. 2017. "Selection against variants in the genome associated with educational attainment." *Proceedings of the National Academy of Sciences of the United States of America* 114 (5): E727–32.

Lumley, Thomas. 2023. "Survey: Analysis of Complex Survey Samples."

Polderman, Tinca J. C., Beben Benyamin, Christiaan A. De Leeuw, Patrick F. Sullivan, Arjen Van Bochoven, Peter M. Visscher, and Danielle Posthuma. 2015. "Meta-analysis of the heritability of human traits based on fifty years of twin studies." *Nature Genetics* 47 (7): 702–9. https://doi.org/10.1038/ng.3285.

Ware, E. B., A. M. Gard, L. L. Schmitz, and J. D. Faul. 2020. "HRS Polygenic Scores – Release 4." Ann Arbor, Michigan: University of Michigan; Survey Research Center, Institute for Social Research.

Wurm, Lee H, and Sebastiano A Fisicaro. 2014. "What Residualizing Predictors in Regres-

sion Analyses Does (and What It Does Not Do)." *Language* 72: 37–48.

Yang, Yang, and S Philip Morgan. 2003. "How Big Are Educational and Racial Fertility Differentials in the US?" *Social Biology* 50 (3-4): 167–87.