Negative Selection in the 20th Century

David Hugh-Jones, Abdel Abdellaoui

April 2020

1 Introduction

Existing work suggests that natural selection is taking place in modern populations, but that effect sizes are small. In this paper we investigate natural selection in the UK over the twentieth century, using data from UK Biobank. We use questions on number of children and number of siblings to reconstruct two generations of family size per respondent. We then relate these to polygenic scores for several different characteristics.

Natural selection is taking place on most polygenic scores, and is broadly "negative" in that characteristics that most people would consider undesirable are being selected for. Effect sizes are larger for number of siblings than for number of children. This is not because the strength of natural selection has changed over time: effect sizes do not differ between early- and late-born respondents, within either children or siblings. Instead, the difference is probably due to ascertainment bias in UK Biobank. Indeed, natural selection is concentrated among population subgroups that are under-represented in the sample, including less-educated people and poorer households. When we weight the sample to correct for ascertainment bias, many effect sizes are considerably larger.

Among mothers who were older at first birth, and among parents with fewer sexual partners, effect signs are reversed so that natural selection is broadly "positive". The effects of PGS can be decomposed into two channels: an effect on age at first live birth, and an opposite-signed effect on number of children conditional on age at first live birth. This pattern may be explained by economic theories of fertility. In these, higher potential earnings have two opposite effects on fertility: a fertility-increasing income effect (higher income makes children more affordable), and a fertility-lowering substitution effect (time spent on childrearing has a higher cost in foregone earnings). Among couples who can take advantage of coordination in childcare activities, the income effect dominates; among single parents or those in unstable relationships, the substitution effect dominates.

2 Open questions/TODO

- If this theory is true then the pattern of differences is explained by (potential) earnings. So the strength of the relationship of a PGS with number of children should be explained by the strength of its relationship with earnings and education.
- Even after weighting, effect sizes for siblings are still higher than for children. Why?
    - Remaining ascertainment bias?
    - Siblings don't include those with 0 children?
    - Something else?
- If we weight with the GHS, and plot the means of figure 1 using this w weighting, most things stay the same - but the change in EA is reversed! EA increases over time. Could this be because the sample includes many relatively educated older people? If so, then mean effects are *overestimated* by figure 1.
- Could be useful: f.2139 - age first had sex (includes "never had sex" which may explain some of the many NAs for f.2141, num sex partners)

- Need f.6138 to correctly calculate `age_fulltime_edu` including those who went to college.
- Look at geography, esp. in siblings regressions.
- Work more on weighting the data
  - check for existing work by others?
  - deal with the fact that GHS 06 has its own weights and use subset.svydesign to reweight?
- control for age in n_partners regressions

3 Data

Data is taken from UK Biobank. Polygenic scores were normalized to mean 0, variance 1.

4 Results

Figure 1 plots mean polygenic scores by 5-year birth intervals, for the entire sample. Several scores show consistent increases or declines over this 30-year period, of the order of 5% of a standard deviation.

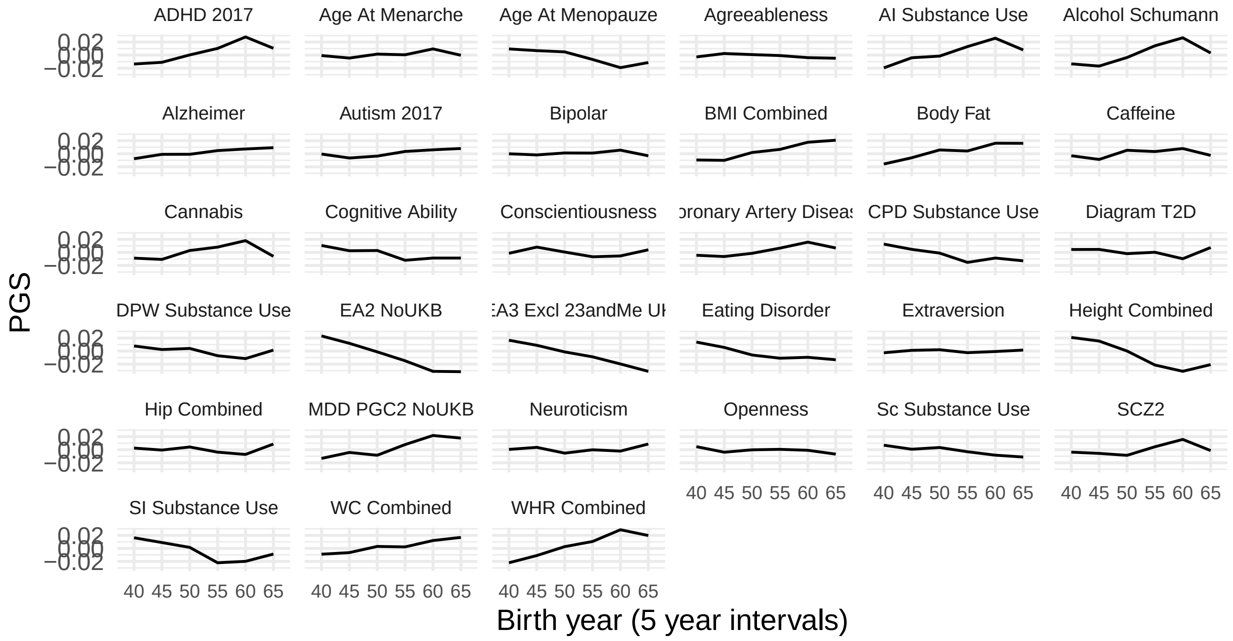


Figure 1: Mean polygenic scores by birth year in UK Biobank. Lines are means for 5-year intervals

We run bivariate regressions on two dependent variables:

- *siblings*, the number of full siblings in the respondent's sib (including himself or herself).
- The number of *children* ever born to/fathered by the respondent.

Figure 2 shows effect sizes of a one-standard deviation shift in each polygenic score.

Estimates are broadly consistent across generations. However, effect sizes are much larger for *siblings* than for *children* regressions. Among consistently-signed estimates, the effect size for siblings is 371.12 per cent at the median.

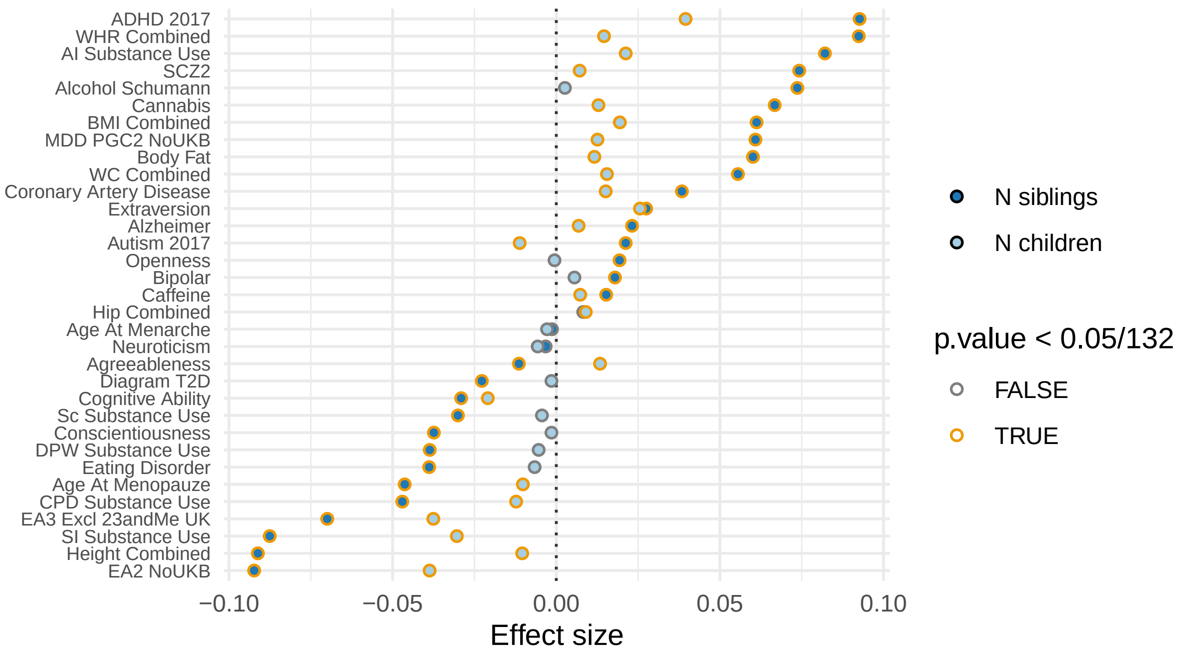


Figure 2: Effects of polygenic scores on number of siblings/children.

5 Subgroups

We next examine how different subgroups of the population contribute to natural selection. These analyses have two goals: to get a sense of how ascertainment bias could affect our basic results, and to learn more about the causes of natural selection in the UK population. Because we have more data on respondents than on their parents, we focus on *children* regressions.

5.1 Sex, income and education

Figure 3 shows effect sizes of PGS on number of children separately for males and females. Differences are particularly large for educational attainment, height and MDD. There is no overall pattern in differences of effect size, however.

Figure 4 splits respondents up by education levels. Selection effects are typically larger and more significant for those who left school before 16. Figure 5 splits respondents by household income category. A very similar pattern holds, with selection effects being larger for those in the poorest income category.

These results could be driven by age, if older respondents are poorer and less educated, and also more subject to selection on polygenic scores. However, if we rerun the regressions, interacting the polygenic score with income category and also with a quadratic in age, the interaction with income remains significant at 0.05/33 for 21 out of 33 regressions. Similarly if we interact the PGS with age of leaving full time education and a quadratic in age, the interaction with age leaving FTE remains significant at 0.05/33 for 12 out of 33 regressions.

The UK Biobank sample overrepresents highly educated people and high-income households. This suggests that statistics from the sample may underestimate the level of natural selection in the population.

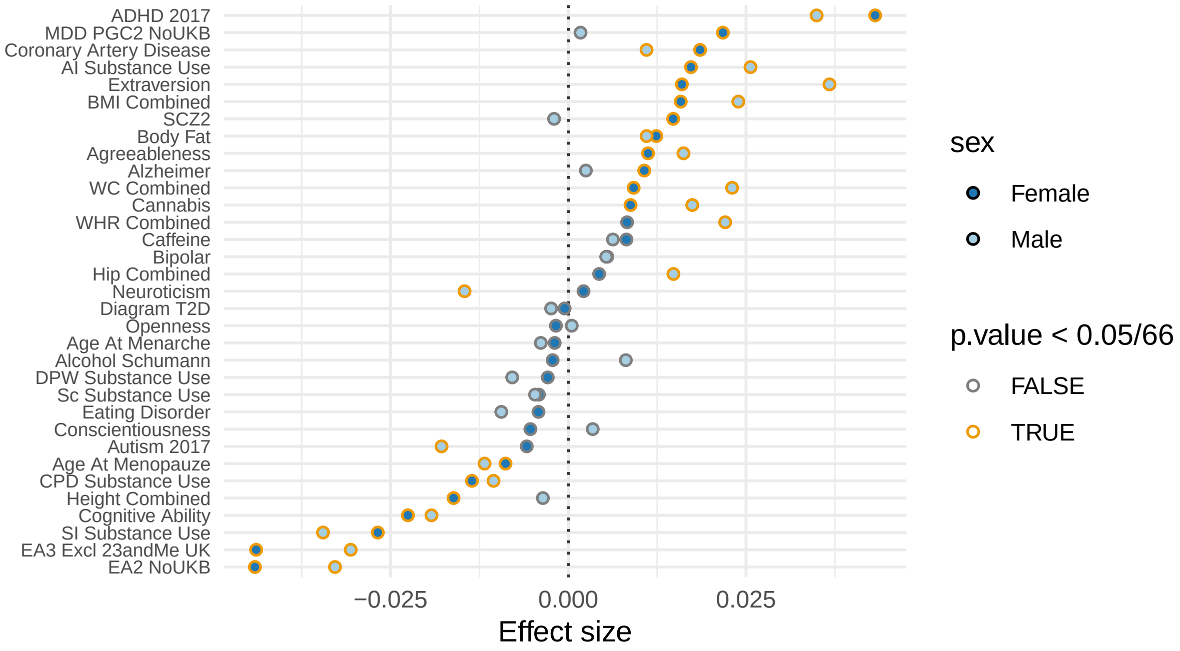


Figure 3: Effect sizes on number of children by sex

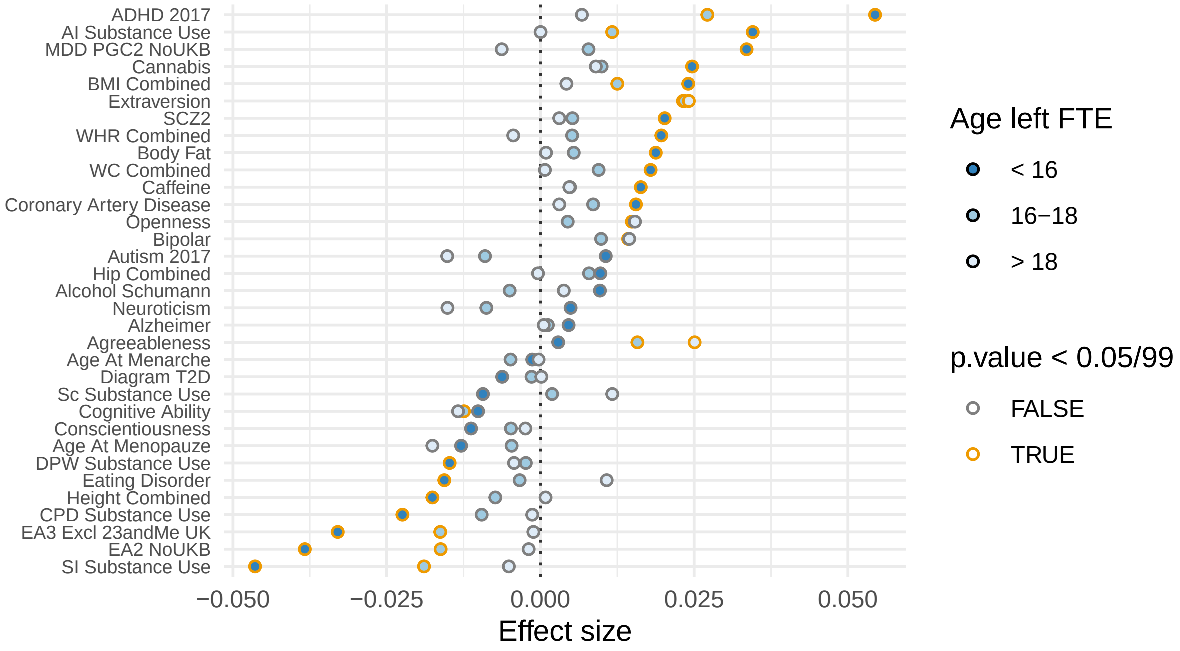


Figure 4: Effect sizes on number of children by age left full-time education

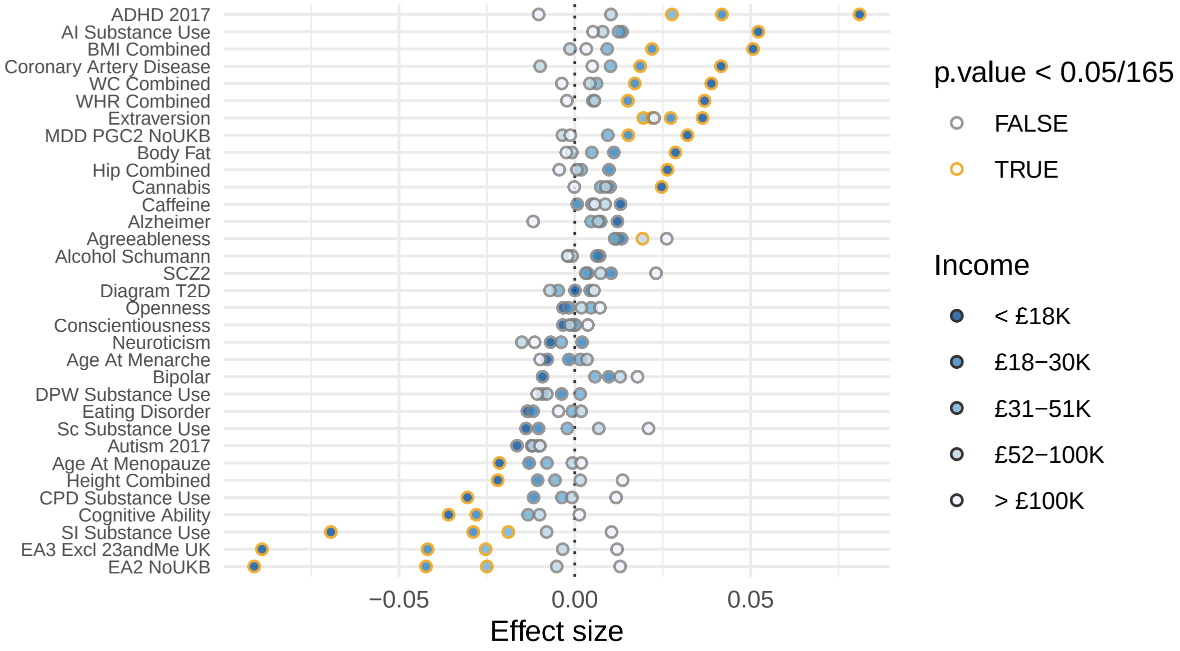


Figure 5: Effect sizes on number of children by household income

5.2 Age at first live birth

Age at first live birth correlates with number of children ever born. Figure 6 shows *children* regressions estimated separately for each tercile of age at first live birth. Several effects are strikingly different across terciles. ADHD and MDD are selected for amongst the youngest third of mothers, but selected against among the oldest two-thirds. Educational attainment is selected for among the oldest two-thirds of mothers, but is not significantly selected among the youngest third. Similarly, several PGS of body measurements are selected against only among older mothers. The correlation between effect sizes for the youngest and oldest terciles is -0.54. This is the first sign that selection may work in different directions across different subgroups.

Figure 7 splits males and females by lifetime number of sexual partners, at the median value of 3. Again, selection effects are reversed across the groups. Among men and women with more than 3 sexual partners, for example, EA is selected against; among those with 3 or fewer sexual partners, it is selected for. Selection is broadly negative (positive) among those with more (less) partners.

6 Accounting for ascertainment bias

As a fix for ascertainment bias in the UK Biobank sample, we weight participants by sex, age (within the 40-71 age group), and education. Our sample weights are calibrated based on the 2006 General Household Survey. We then rerun the basic *children* regressions. Figure 8 shows the results, with unweighted estimates plotted for comparison. Weighting increases effect sizes on average by 26.18 per cent.

We also weighted data for women with children only, by age, education and age at first live birth. Figure 9 shows the results. Weighting increased effect sizes on average by -5.24 per cent. There are large changes for some variables, including EA3 (103.77 per cent), cognitive ability (102.98 per cent) and waist-hip ratio (WHR; 146.76 percent).

These results suggest that previous work may have underestimated the level of selection occurring in the population. [XXX who?]

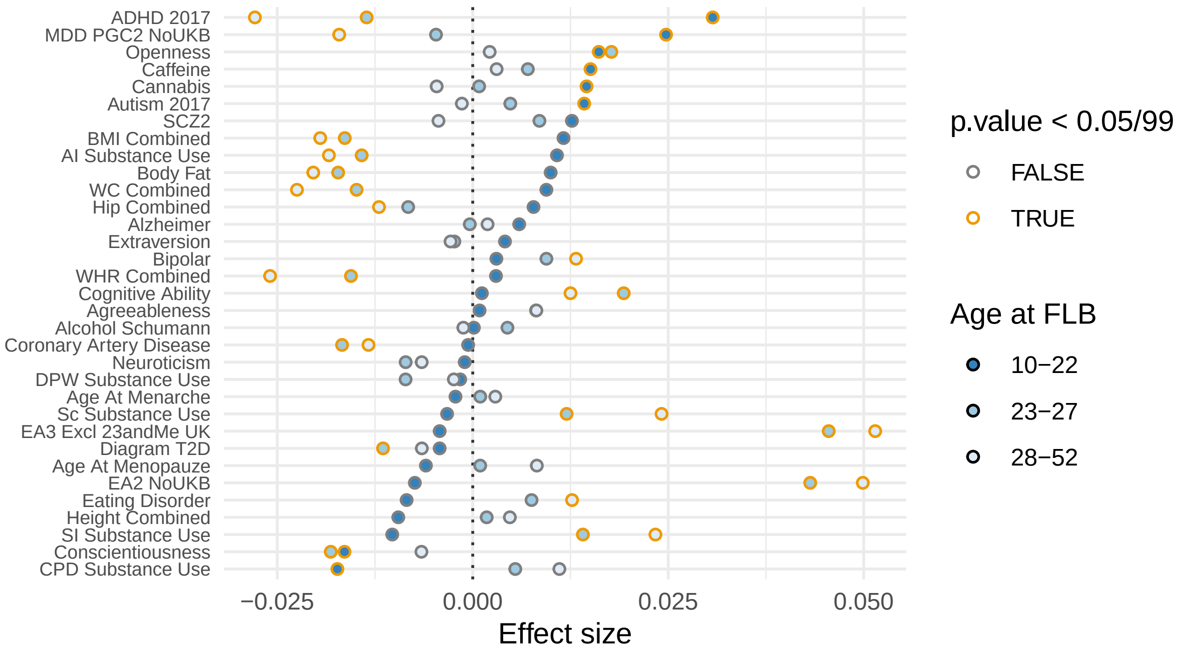


Figure 6: Effect sizes on number of children, by age at first live birth terciles

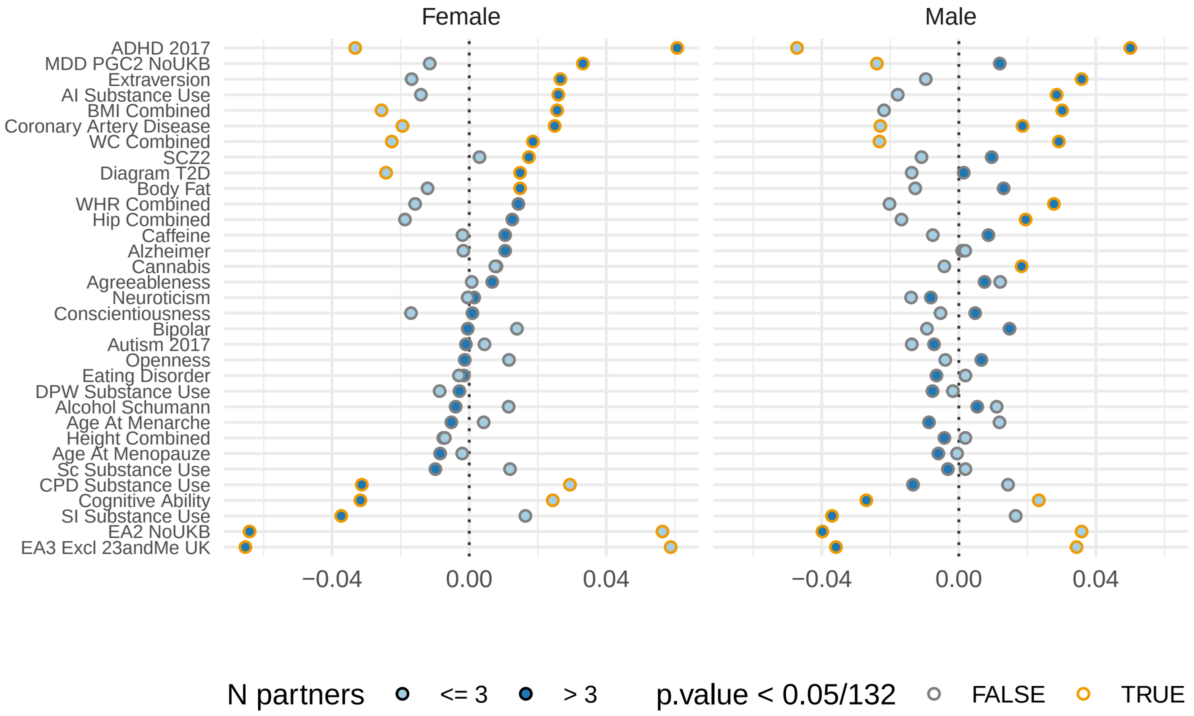


Figure 7: Effect sizes on number of children by lnumber of sexual partners

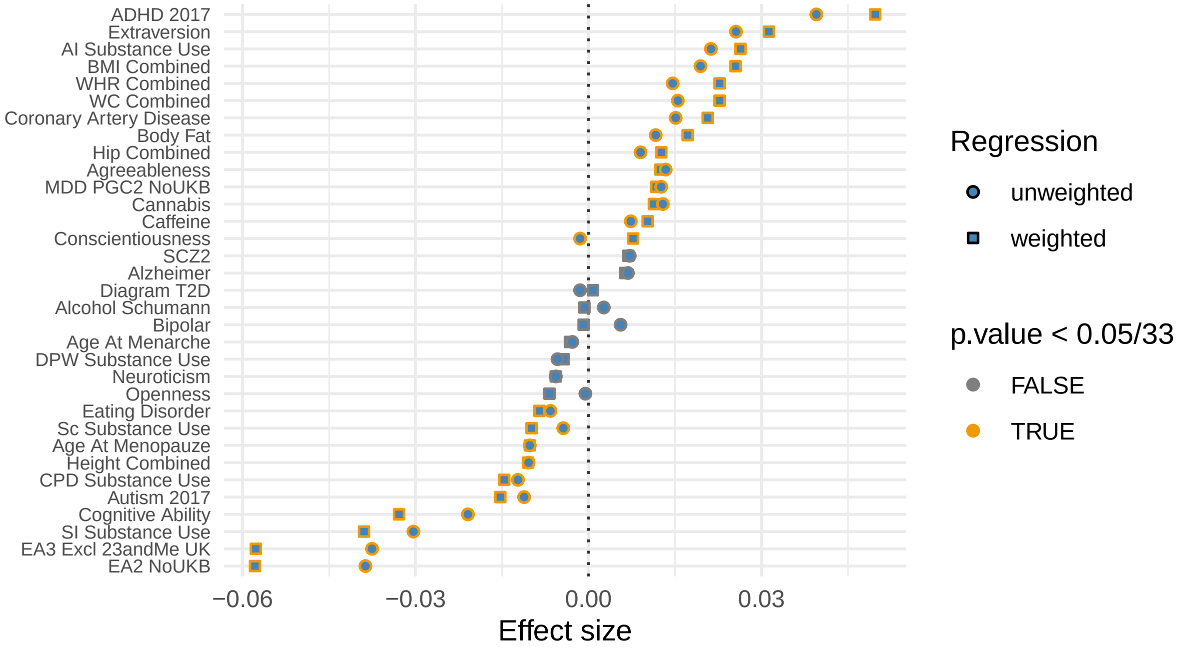


Figure 8: Effects of polygenic scores on number of children, regressions weighted by education levels within age categories

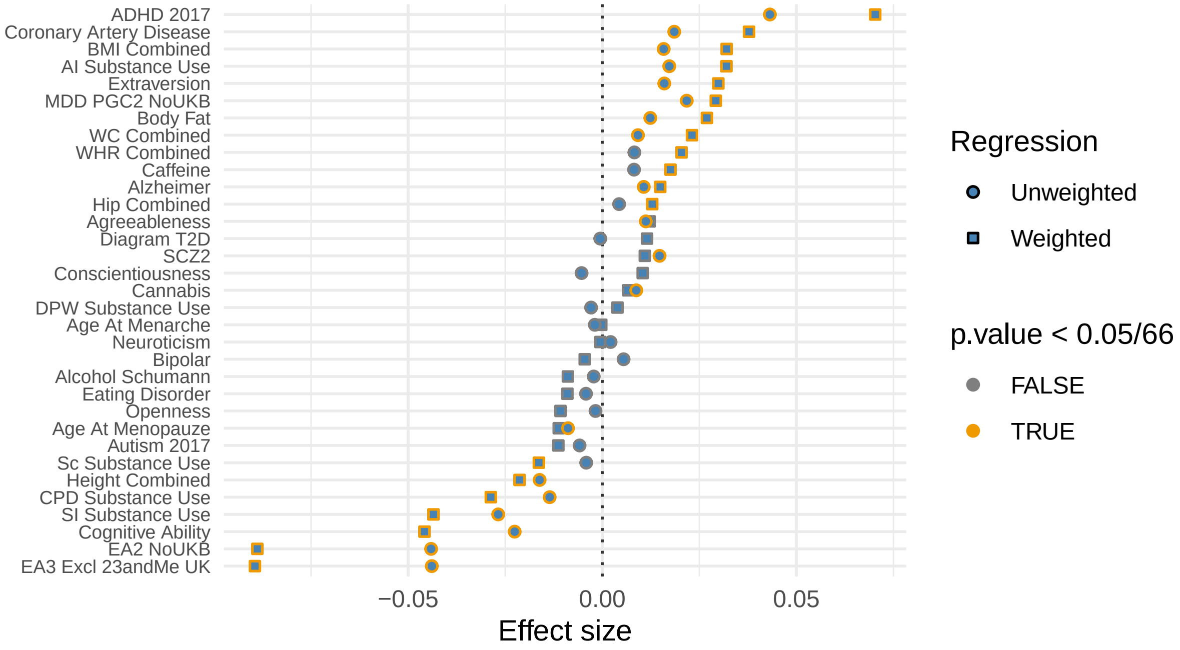


Figure 9: Effect sizes on number of children, female respondents weighted by age at first live birth

6.1 Mechanisms behind natural selection

In this section, we dig into the mechanisms causing natural selection in the population. In particular, why are selection effects *oppositely* signed in certain subgroups – for example, in older and younger mothers (by age at first live birth) or in those with more and fewer lifetime sexual partners?

A possible answer is given by the economic theory of fertility [XXX cite Becker 1960]. According to this, increases in potential earned income affect fertility via two opposing channels. There is an "income effect" by which children become more affordable, like any other good. Since childrearing has a cost in time, there is also a substitution effect: the opportunity cost of childrearing increases if one's market wage is higher. The income effect would lead higher earners to have more children. The substitution effect would lead higher earners to have fewer children.

Traditionally, it was assumed that since women spent more of their time on childcare, the substitution effect would dominate for women's earnings, while the income effect would dominate for men's earnings. However, a more relevant split may be between those raising children alone, and those raising children in a two-parent household. The substitution effect is likely to be particularly strong for lone parents, since they have less opportunity to share childcare responsibilities.

If so, then genetic characteristics which affect one's earnings potential in the labour market may lead to opposing effects on fertility. Genetic variants which improve one's earnings may increase fertility among couples, but decrease fertility among single parents.

This explanation predicts that the strength of polygenic scores' effects on fertility will correlate with the strength of their relationship with earnings. We measure the correlation of each PGS with household income, as well as with education level, a phenotypic variable which is likely to predict earnings. We then correlate these measurements with effect sizes on number of children. Across scores, there is a clear negative correlation between effect on fertility, and link to both income and education.

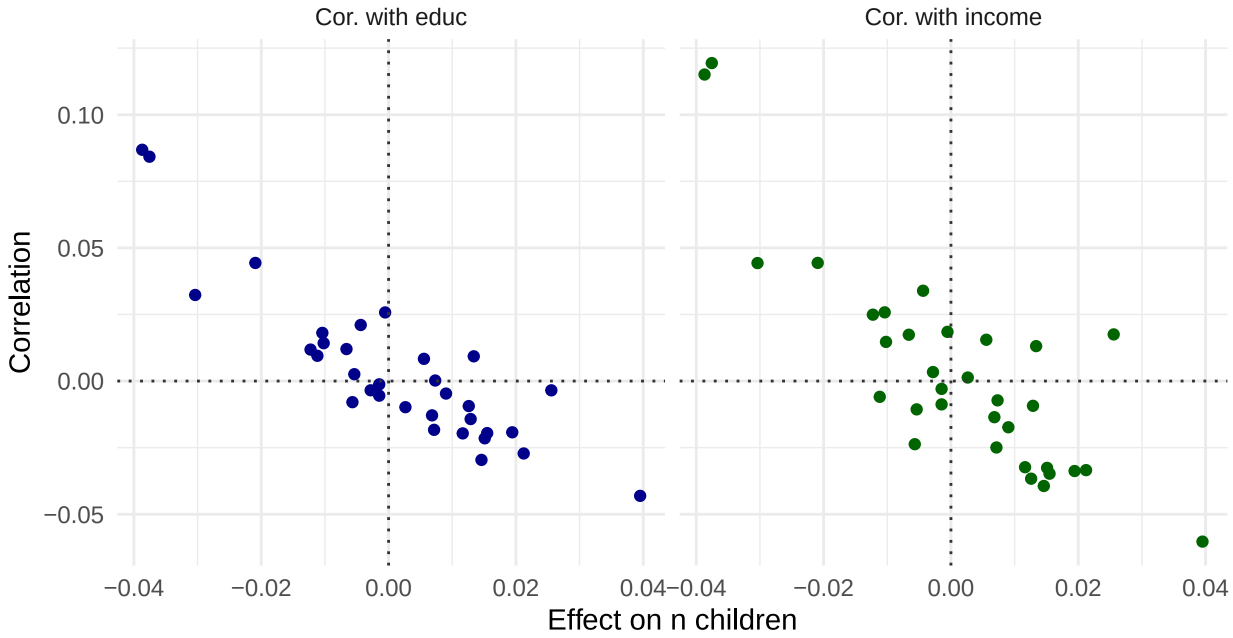
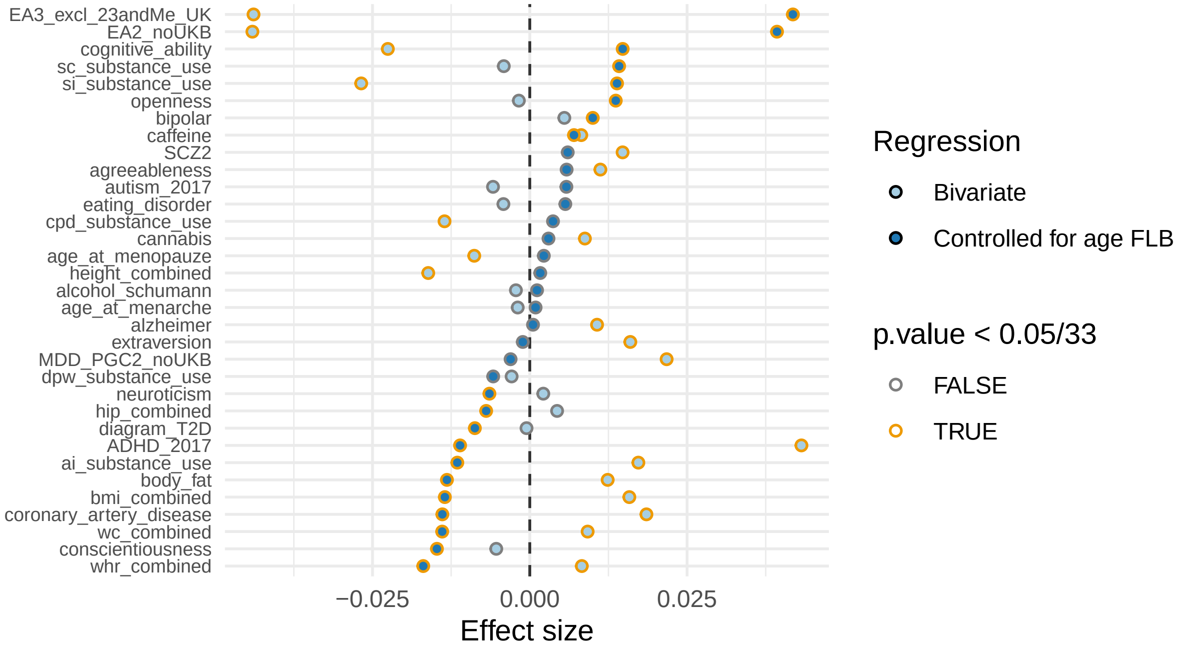


Figure 10: Effect sizes on number of children and correlations with income and education

7 Bits to integrate

Figure **??** shows the results of *children* regressions for women only, controlling for age at first live birth. Effect sizes are greatly reduced. Partly, this is because the regressions exclude childless women. However, in 24 out of 33 cases, effect sizes actually change sign with the controls. The correlation between effect sizes controlling for age at first live birth, and raw effect sizes, is -0.75.



We can run similar regressions for the parents' generation, using the subsets of respondents who reported their mother's or father's age and who had no elder siblings. We run *sibling* regressions on these subsets, controlling for either parent's age at their birth. Figure 11 shows the results. Effect sizes are very similar, whether controlling for father's or mother's age at respondent's birth or mother's age at respondent's birth. Unlike for the respondents' generation, effect sizes are positively correlated with the effect sizes from bivariate regressions (father's age at birth: ρ 0.51; mother's age at birth: ρ 0.57).

These results suggest that polygenic scores may directly correlate with age at first live birth. Figure 12 plots estimated effect sizes from bivariate regressions for respondents, and Figure 13 does the same for their parents. Effect sizes are reasonably large. They are also very highly correlated across generations. Effect sizes of PGS on father's age at own birth, and on own age at first live birth, have a correlation of 0.98; for mother's age and own age it is 0.98

8 Appendix

8.1 Controlling for principal components

Polygenic scores could capture effects that are really due to population stratification, although would not change our results for natural selection of the scores. In 14 we show results for selection on polygenic scores residualized for the top 100 principal components of the genetic data, calculated within the UK Biobank population. (XXX Abdel details.)

In siblings regressions, effect sizes are smaller when residualizing for principal components – sometimes much smaller, as in the case of height. 25 out of 33 "controlled" effect sizes have a smaller absolute value than the corresponding "raw" effect size. The median proportion between raw and controlled effect sizes is 0.89.

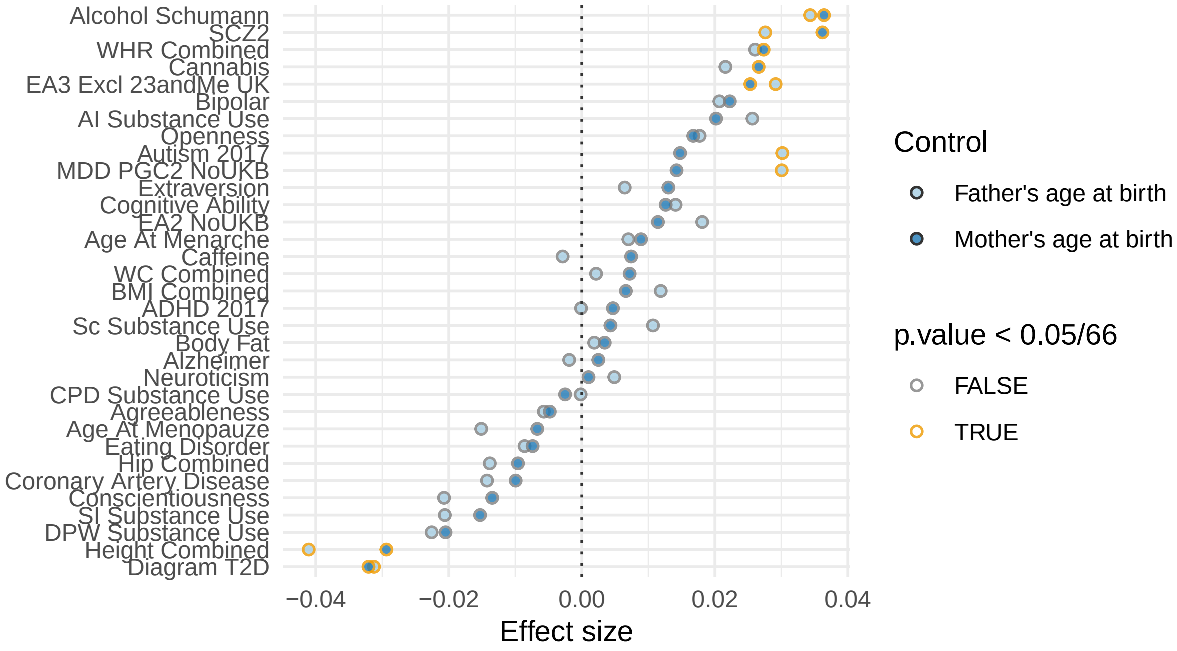


Figure 11: Effect sizes on number of siblings controlling for parents' age at birth, eldest siblings

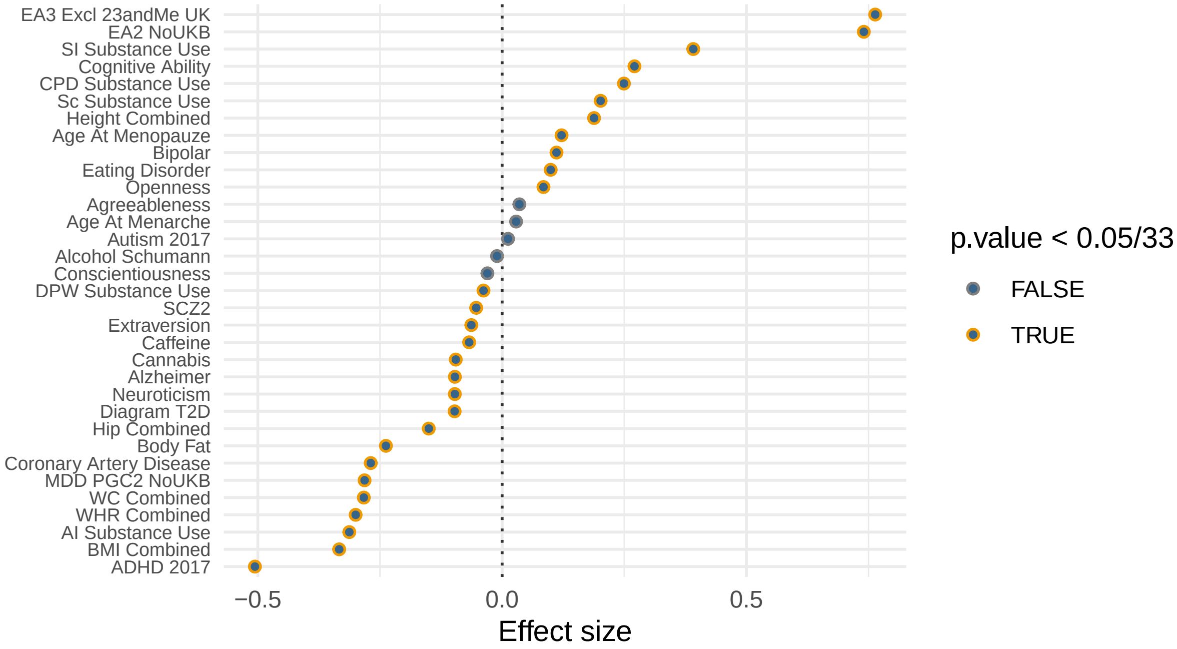


Figure 12: Effects of polygenic scores on age at first live birth

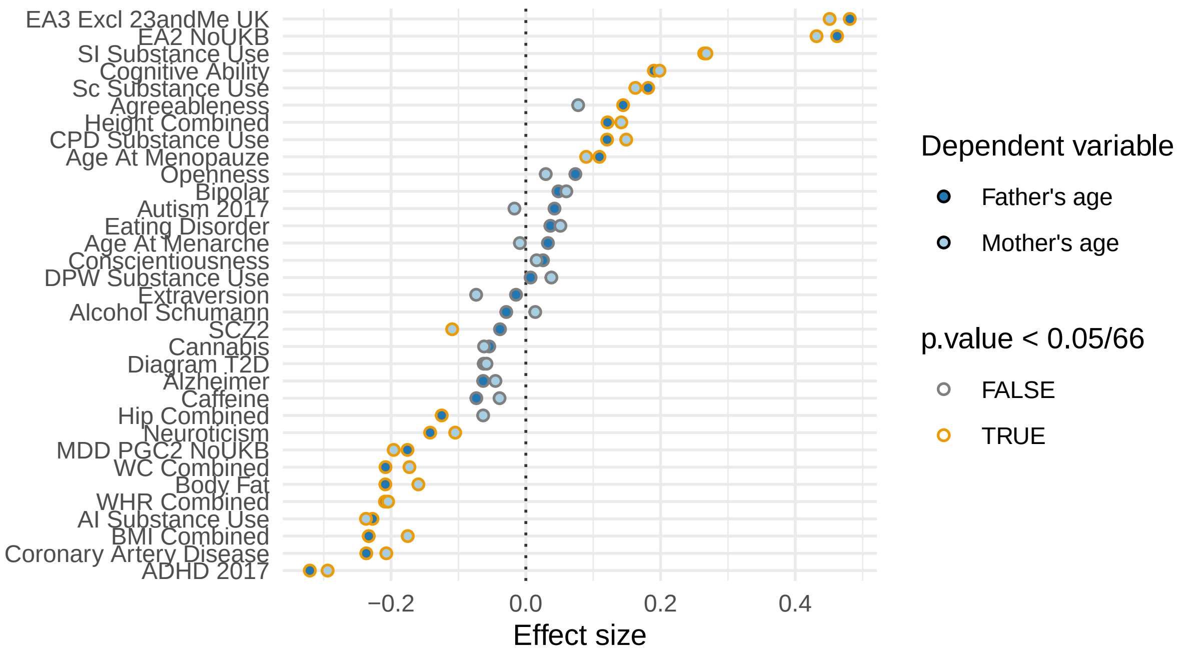


Figure 13: Effects of polygenic scores on parents' age at own birth, eldest siblings

Among the children regressions, this no longer holds. Effect sizes are barely affected by controlling for principal components.

Overall, 78.79 per cent of effect sizes are consistently signed across all four regressions (on children and siblings, and with and without residualization).

To get a further insight into this we regress *siblings* and *children* on individual principal components. As Figure 15 shows, effects are larger and more significant in siblings regressions. 29 principal components significantly predicted number of siblings, while only 10 significantly predicted number of children.

8.2 Selection over time

To check whether effect sizes were changing over time, we ran regressions interacting PGS with birth year, median split at 1950. Tables 1 and 2 summarize the results. Very few scores are significantly different. Only one score, EA3, changes significantly across time in both generations: the absolute size of the (negative) effect significantly increases in sibling regressions, and significantly decreases in children regressions.

Table 1: Change in effect sizes between early and late born parents, 'sibling' regressions

Change	Number of scores
Insignificant	31
Size decreasing	2

Significance is measured at p < 0.05/66

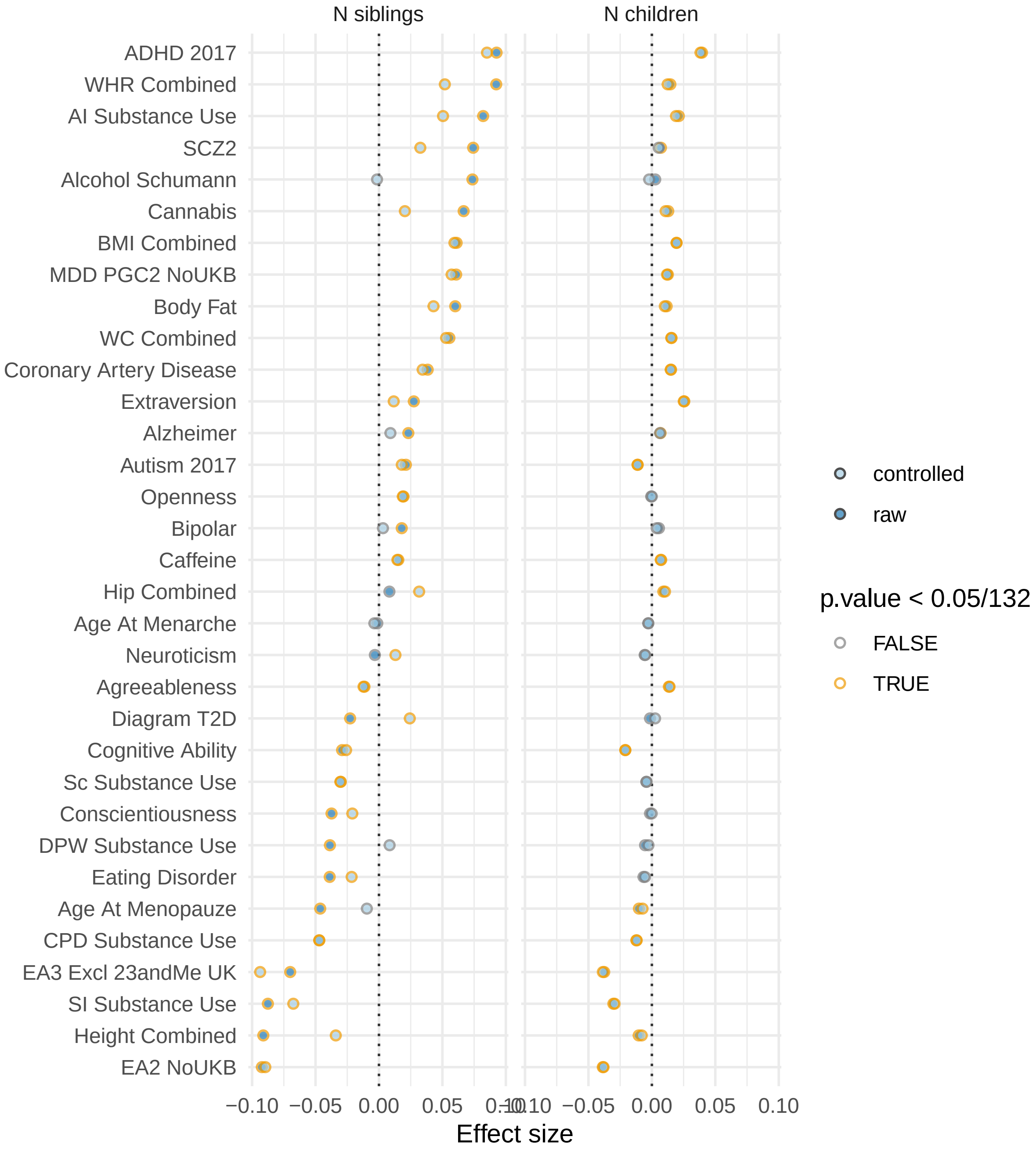


Figure 14: Effects of residualized polygenic scores on number of siblings/children.

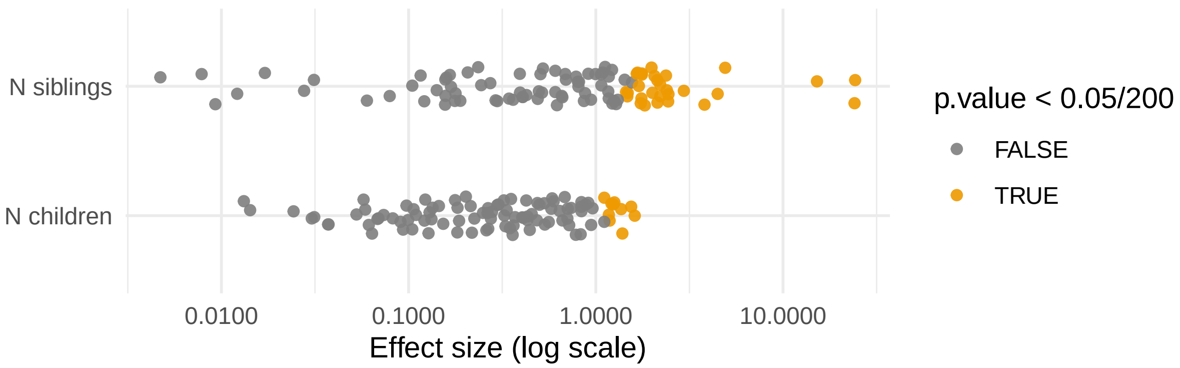


Figure 15: Effect of principal components of genetic data on number of siblings/children. Absolute effect sizes are plotted. Each dot represents one bivariate regression. Points are jittered on the Y axis.

Table 2: Change in effect sizes between early and late born respondents, 'children' regressions

Change	Number of scores
Change sign	2
Insignificant	29
Size increasing	2

Significance is measured at p < 0.05/66

8.3 Causality

8.4 Number of children

Figure 16 shows the full distribution of number of children born for different ventiles of the EA3 polygenic score. The strongest relationship seems to be for having 0 children versus 1 or more.

Table 3: Top 10 correlations between polygenic scores

PGS	PGS	Correlation
EA2 NoUKB	EA3 Excl 23andMe UK	0.89
Hip Combined	WC Combined	0.807
BMI Combined	WC Combined	0.753
WC Combined	WHR Combined	0.711
BMI Combined	Hip Combined	0.697
Body Fat	WC Combined	0.435
BMI Combined	Body Fat	0.425
BMI Combined	WHR Combined	0.425
Body Fat	Hip Combined	0.385
ADHD 2017	Autism 2017	0.328

Different polygenic scores are correlated. Table 3 shows the top correlations in the sample. Because of this, bivariate correlations between PGS and number of children might be driven by other genetic scores. To explore which polygenic scores are driving negative selection, we run a single omnibus regression of *number*

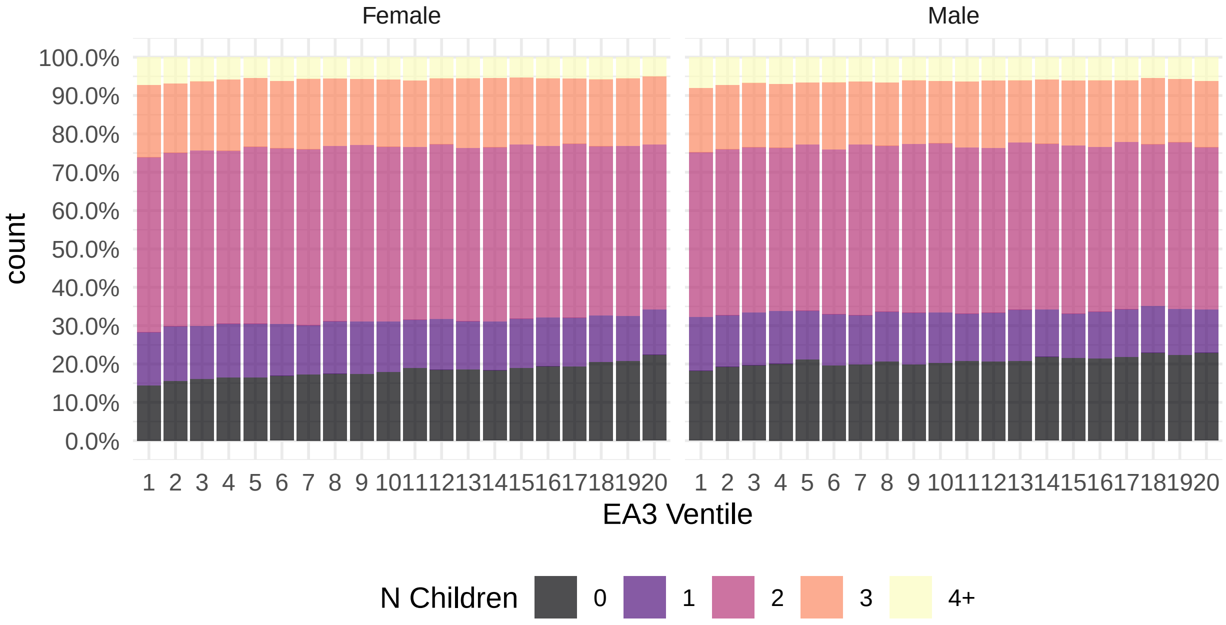


Figure 16: Number of children by ventiles of EA3 PGS

of children on all the PGS. We exclude EA2, waist-hip ratio, waist-circumference, and “Hip combined” since they are highly correlated with other scores, which could make our estimates unstable. Figure 17 shows the results. Interestingly, several PGS remain independently significant, although effect sizes are reduced.

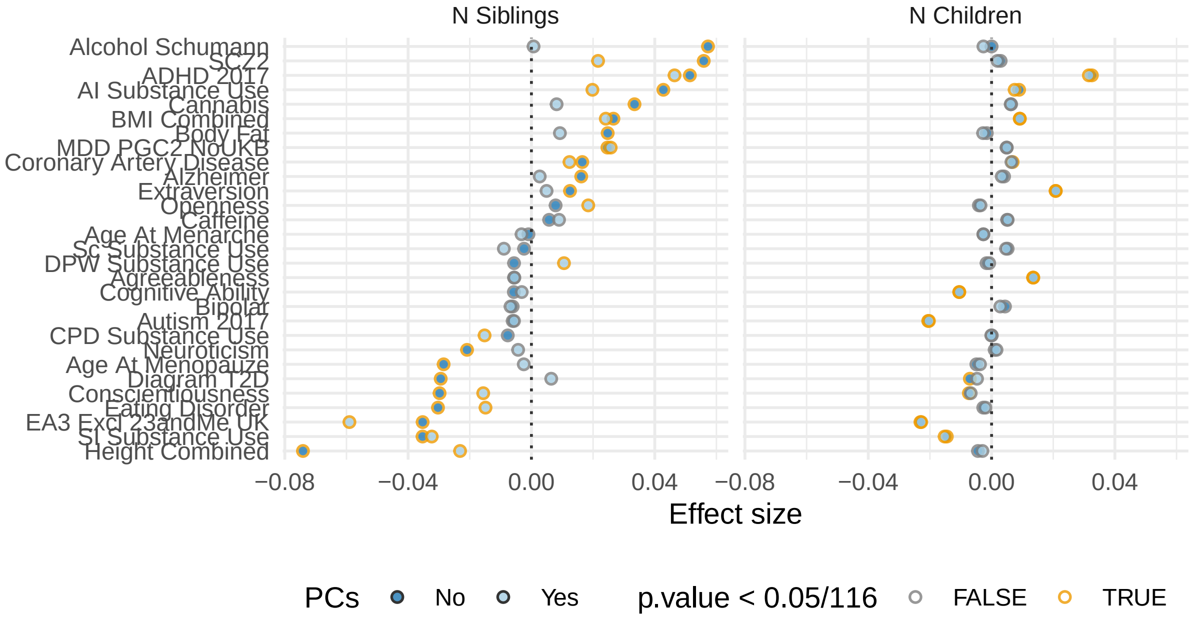


Figure 17: Partial correlations with number of children