Negative Selection in the 20th Century

David Hugh-Jones, Abdel Abdellaoui

April 2020

```
# TODO

* Why so many missings for age_fulltime_edu?
  - ah, need f.6138 for those who went to college
  - but note that YearsEdu.ISCED has many fewer missings
  - ask Abdel how it was calculated
  - use this for survey weighting?

* Look at geography, esp. in siblings regressions.

* Divide siblings regressions by own YOB, into early and late.
  - Use this to confirm the basic "selection decreasing" story. DONE.
  - Actually almost no change.
  - Overall, diffs between siblings and children regressions
    now look more like ascertainment bias than change over time.
  - (I checked my code: I hadn't been controlling for one main term when
    running the interaction. Doh.)

* Work more on weighting the data?
  - weight by age at first live birth
  - education using ISCED

* direct regressions of scores on age_flb

* control for age in n_partners regressions


## Data to gather
* f.2139 - age first had sex (includes "never had sex" which may explain
  some of the many NAs for f.2141, num sex partners)
```

1 Data

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

2 Results

```
## `summarise()` ungrouping (override with `.groups` argument)
```

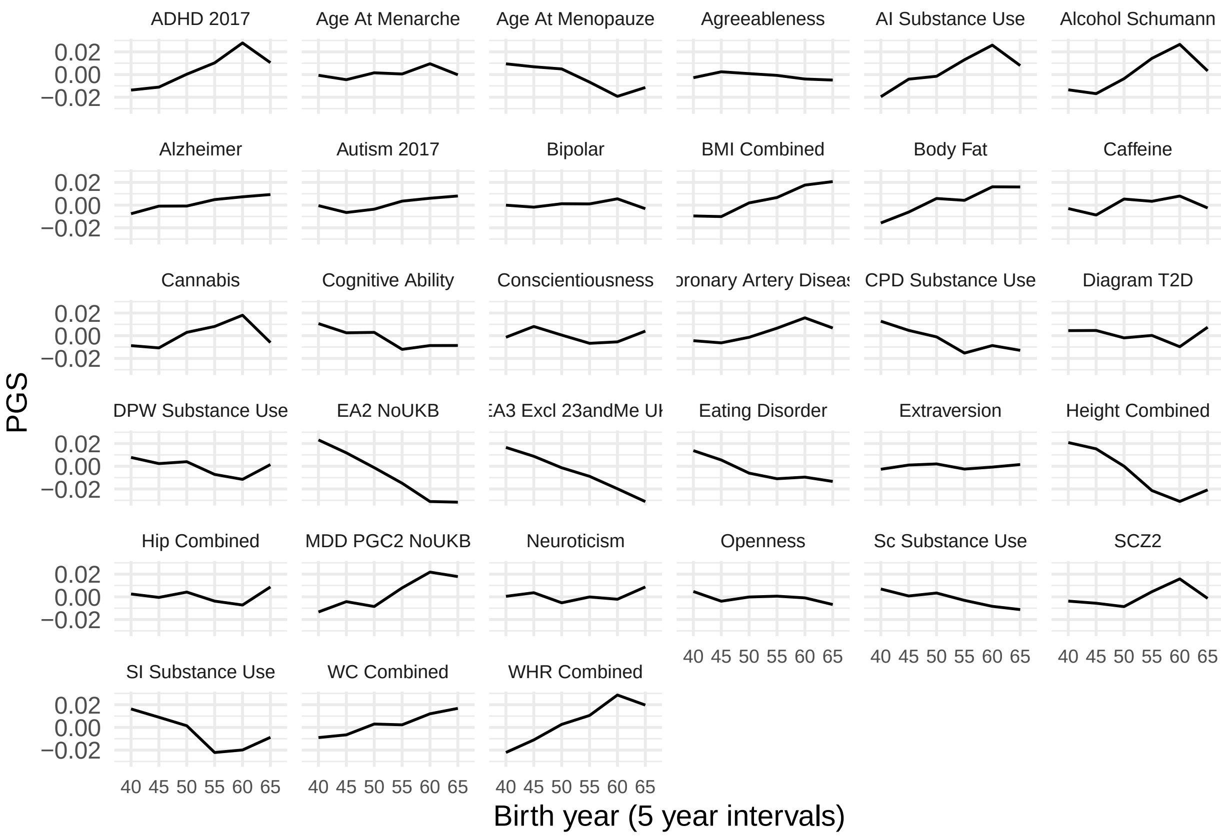


Figure 1: Mean polygenic scores by birth year in UK Biobank. Lines are means for 5-year intervals

We run 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.

We run regressions both with and without controls for the 100 top principal components of the genetic data.

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

Estimates are broadly consistent across generations. For 27 out of 33 polygenic scores, all 4 estimates have the same sign.

However, effect sizes are much smaller for *children* than *siblings* regressions. Among consistently-signed estimates, the median effect size for children as a proportion of the effect size for siblings is 0.27, or 0.4 with controls.

In siblings regressions, effect sizes are smaller when controlling for principal components – sometimes much smaller, as in the case of height. 20 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 0.92. Among the children regressions, this no longer holds. Effect sizes are barely affected by controlling for principal components.

```
## `summarise()` ungrouping (override with `.groups` argument)
```

To get a further insight into this we regress *siblings* and *children* on individual principal components. As Figure 3 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.

3 Selection over time

Negative selection seems to decrease over time. Figure 4 shows effect sizes for *number of siblings* and *number of children*, median-split by parents' year of birth and own year of birth respectively. Parents' year of birth is imputed, which is likely to produce some bias.

```
## Scale for 'colour' is already present. Adding another scale for 'colour',
## which will replace the existing scale.
```

By definition, the sibling regressions exclude members of the parents' generation who had no children. This is likely to bias results towards zero, since much of the effect in children regressions is due to respondents with high scores being more likely to have no children. So, we cannot directly compare effect sizes for the two sets of regressions. Within the sibling regressions, the most common pattern is that negative effects shrink in absolute size (Table 1).

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 $p < 0.05/66$

In children regressions, no clear pattern is visible (Table 2).

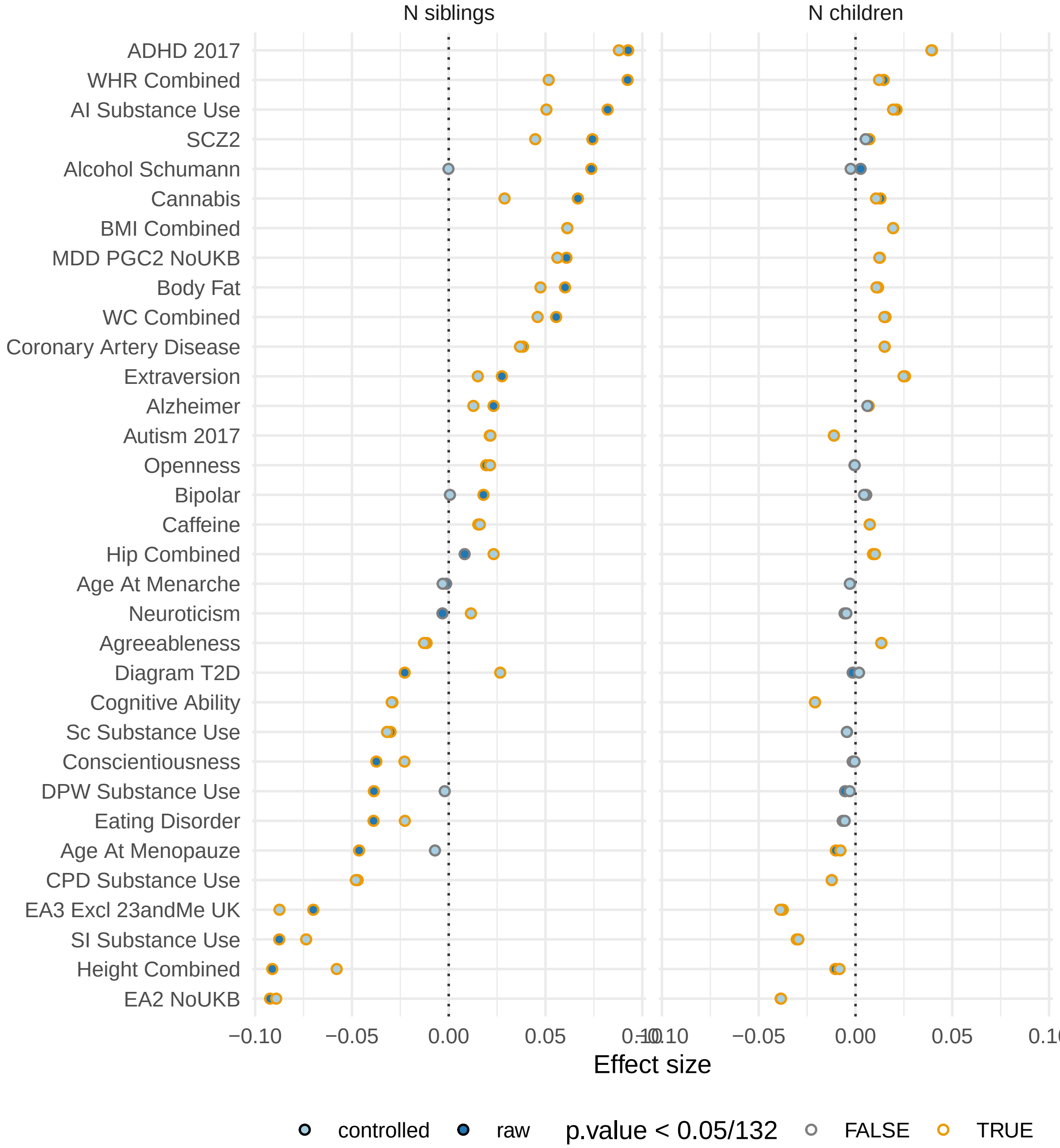


Figure 2: Effects of polygenic scores on number of siblings/children.

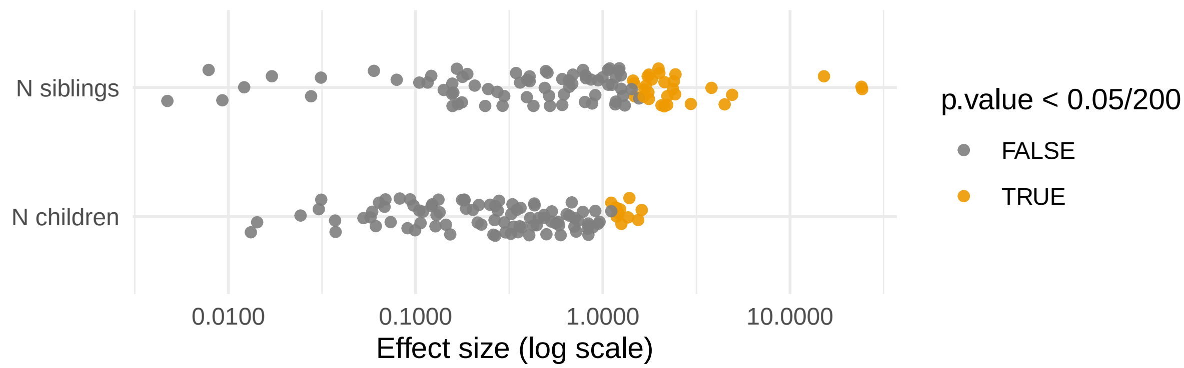


Figure 3: Effect of principal components of genetic data on number of siblings/children. Absolute effect sizes are plotted. Each dot represents one bivariate regression. Points are jittered on the Y axis.

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 $p < 0.05/66$

4 Accounting for ascertainment bias

Effect sizes tend to be smaller for children regressions. This could be caused by ascertainment bias in the UK Biobank sample – e.g., if respondents themselves are a more selected sample than the respondents' parents. To check this, we weighted UK Biobank participants by age of leaving full time education. Our sample weights are based on the 2006 General Household Survey, calculating proportions within age and sex cells among White British respondents. We then rerun the basic *children* regressions.

Figure 5 shows the results, with unweighted estimates plotted for comparison. Weighting increases effect sizes on average by 39.72 per cent. These weights are only basic attempts to correct for ascertainment bias. Effect sizes might be increased further by more precise weighting.

5 Causality

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 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 6 shows the results. Interestingly, several PGS remain independently significant, although effect sizes are reduced.

6 Subgroups

We next examine how different subgroups contribute to natural selection.

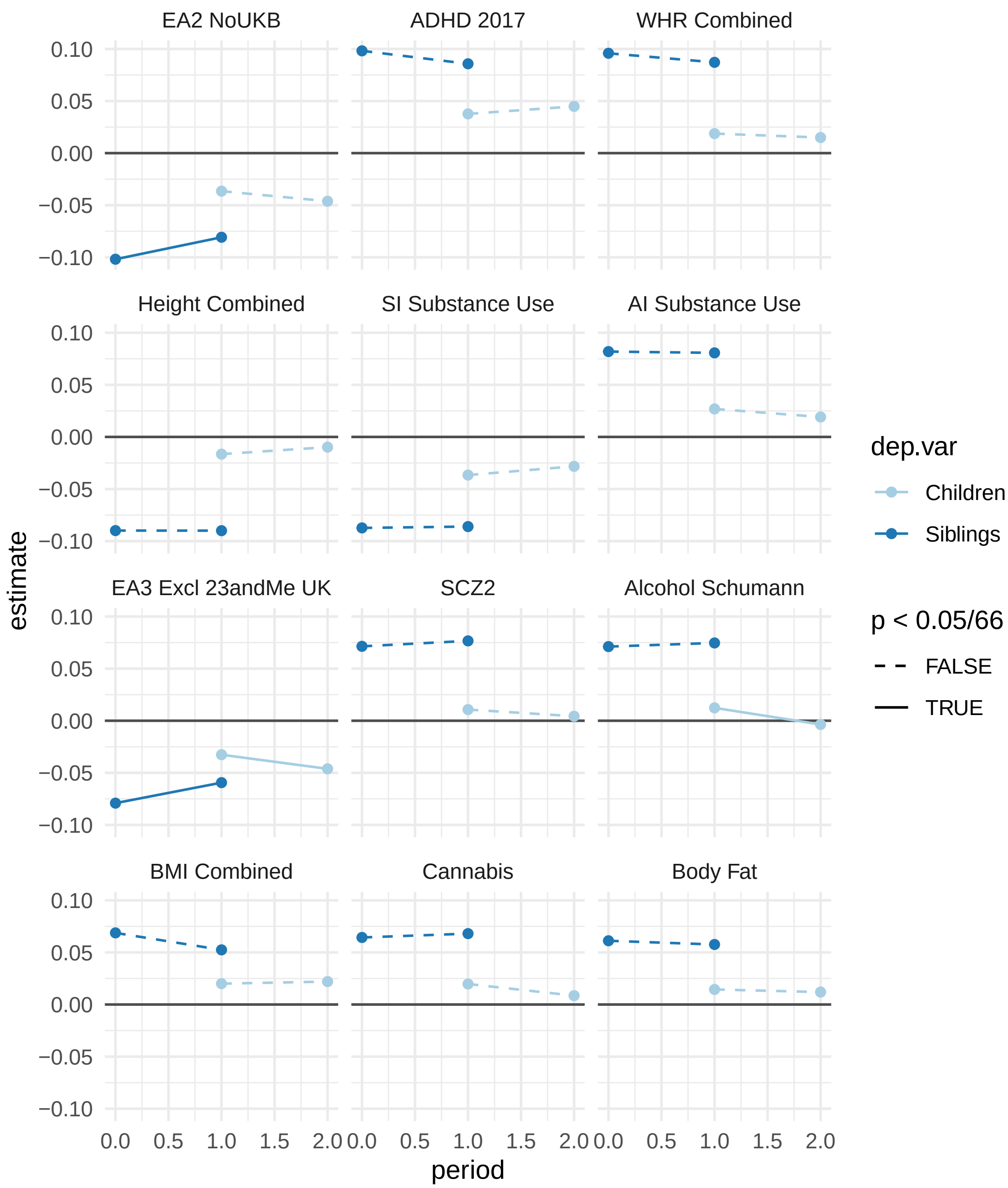


Figure 4: Effect sizes of PGS on number of children/siblings by own/parents' year of birth. PGS with the largest mean effect sizes are shown.

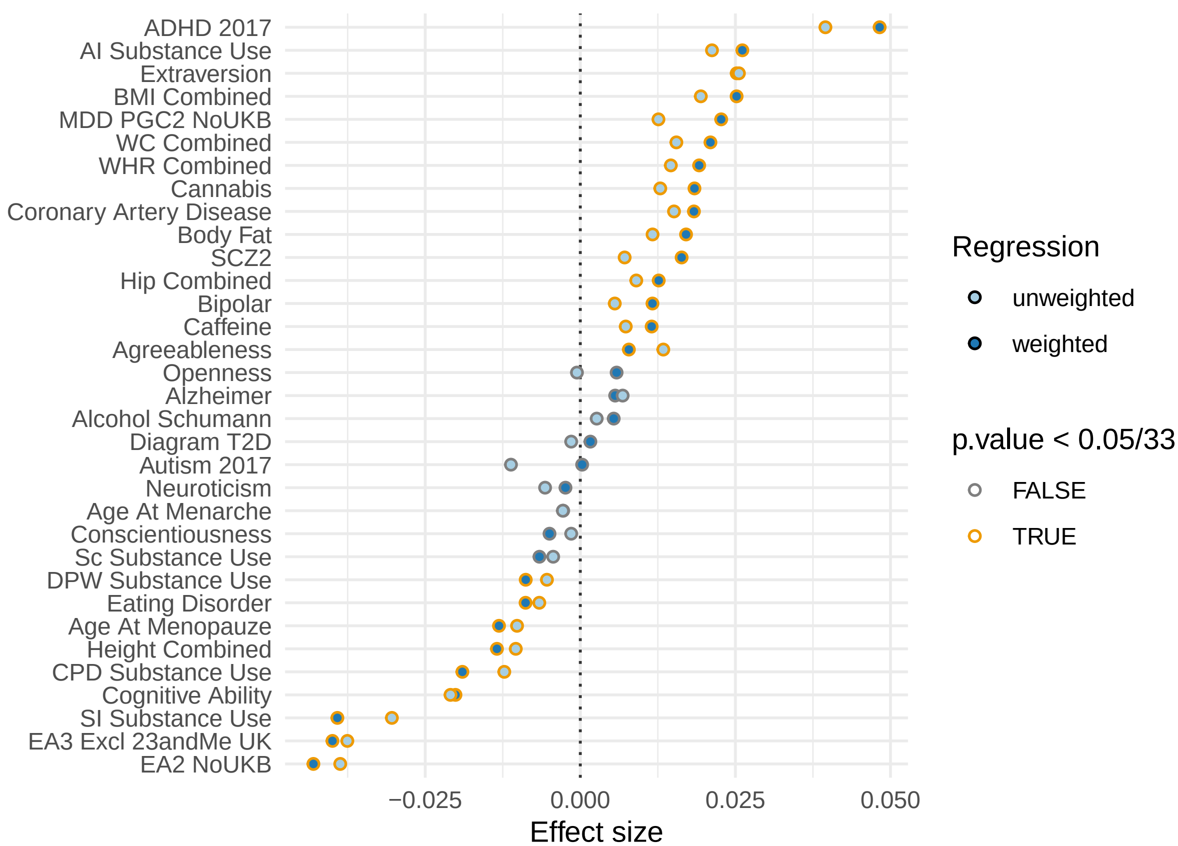


Figure 5: Effects of polygenic scores on number of children, regressions weighted by education levels within age categories

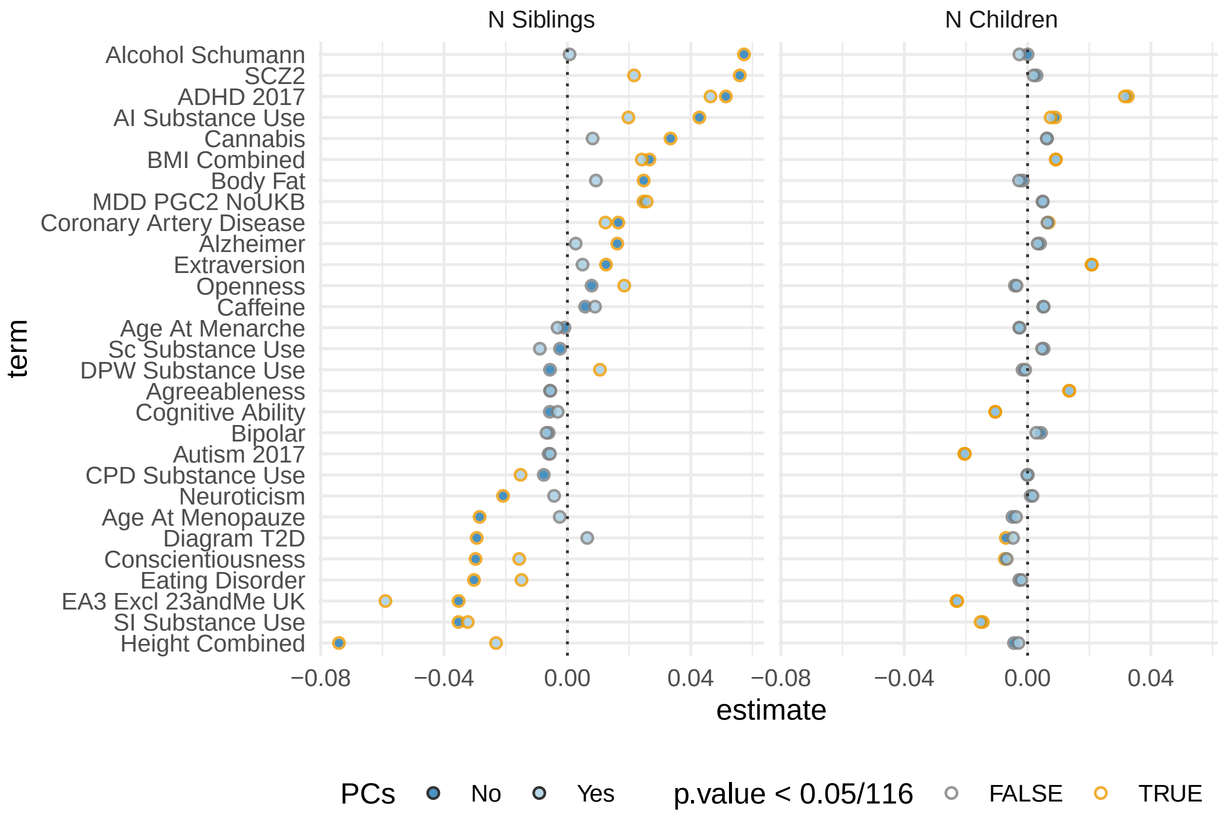


Figure 6: Partial correlations with number of children

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

6.1 Males and females

Figure 7 shows effect sizes of PGS on number of children separately for males and females. For 16 out of 33 PGS, selection is more negative for women than for men. Differences are particularly large for educational attainment and height PGS.

6.2 Age at first live birth

Figure 8 shows the results of *children* regressions for women only, controlling for age at first live birth. Effect sizes are greatly reduced. In 24 out of 33 cases, they are of the opposite sign. The correlation between effect sizes controlling for age at first live birth, and raw effect sizes, is -0.75.

The dataset has no information on fathers' age at birth of their first child. However, we can calculate this information for the parents' generation, for 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 **??** 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 respondent's own 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).

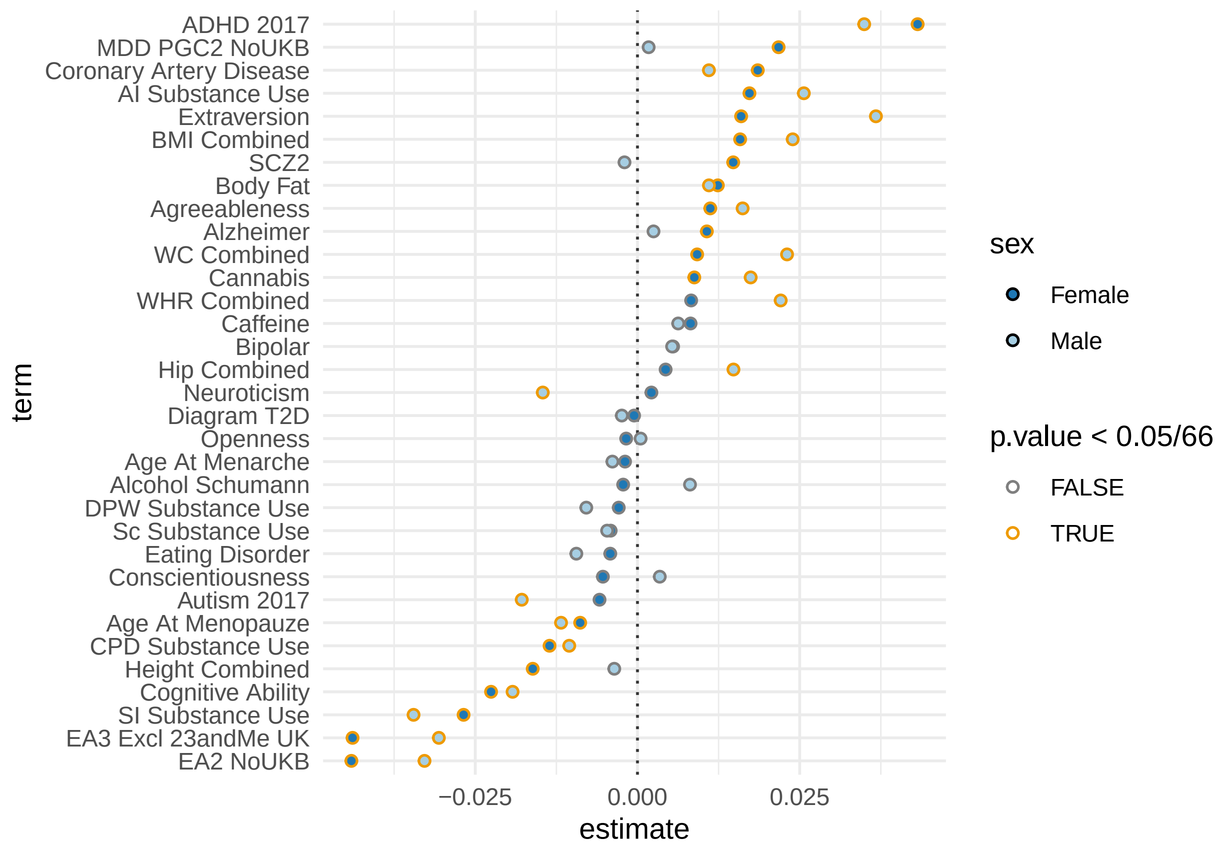


Figure 7: Effect sizes on number of children by sex

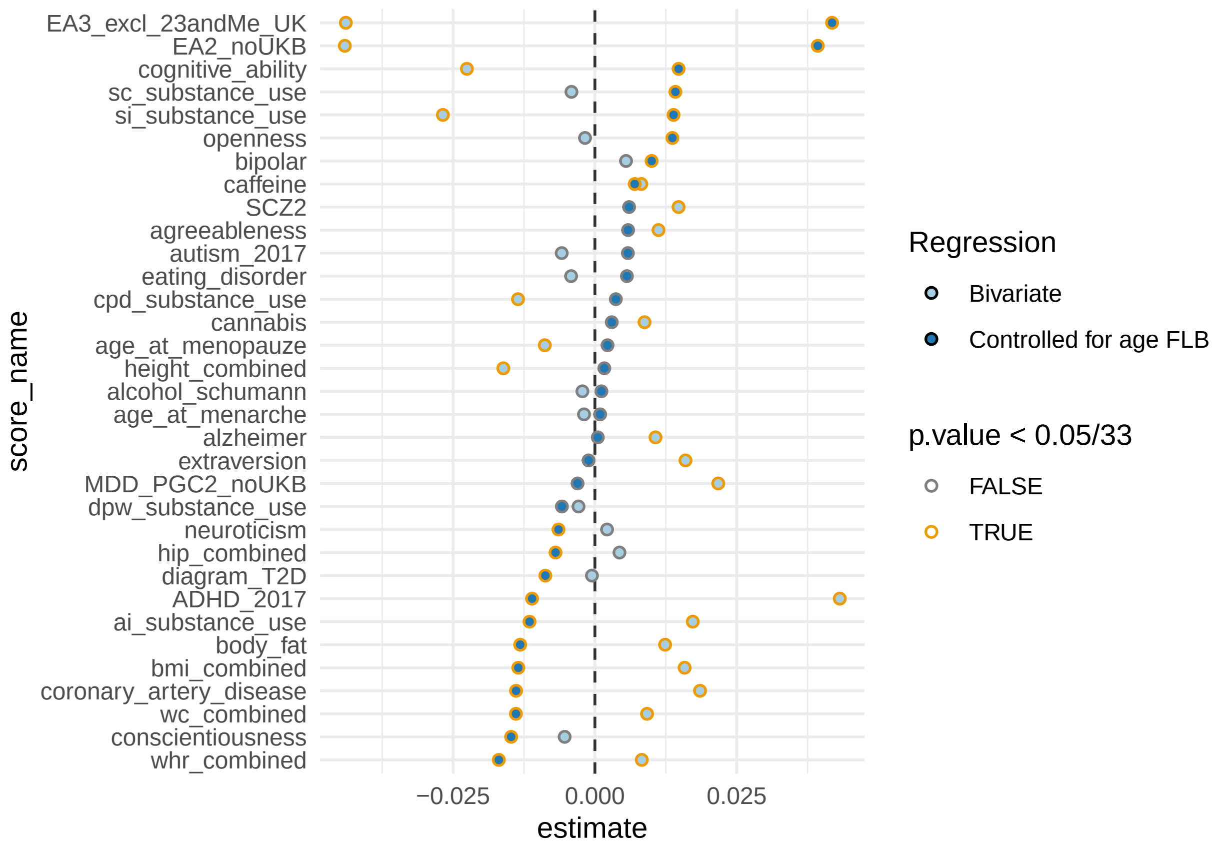
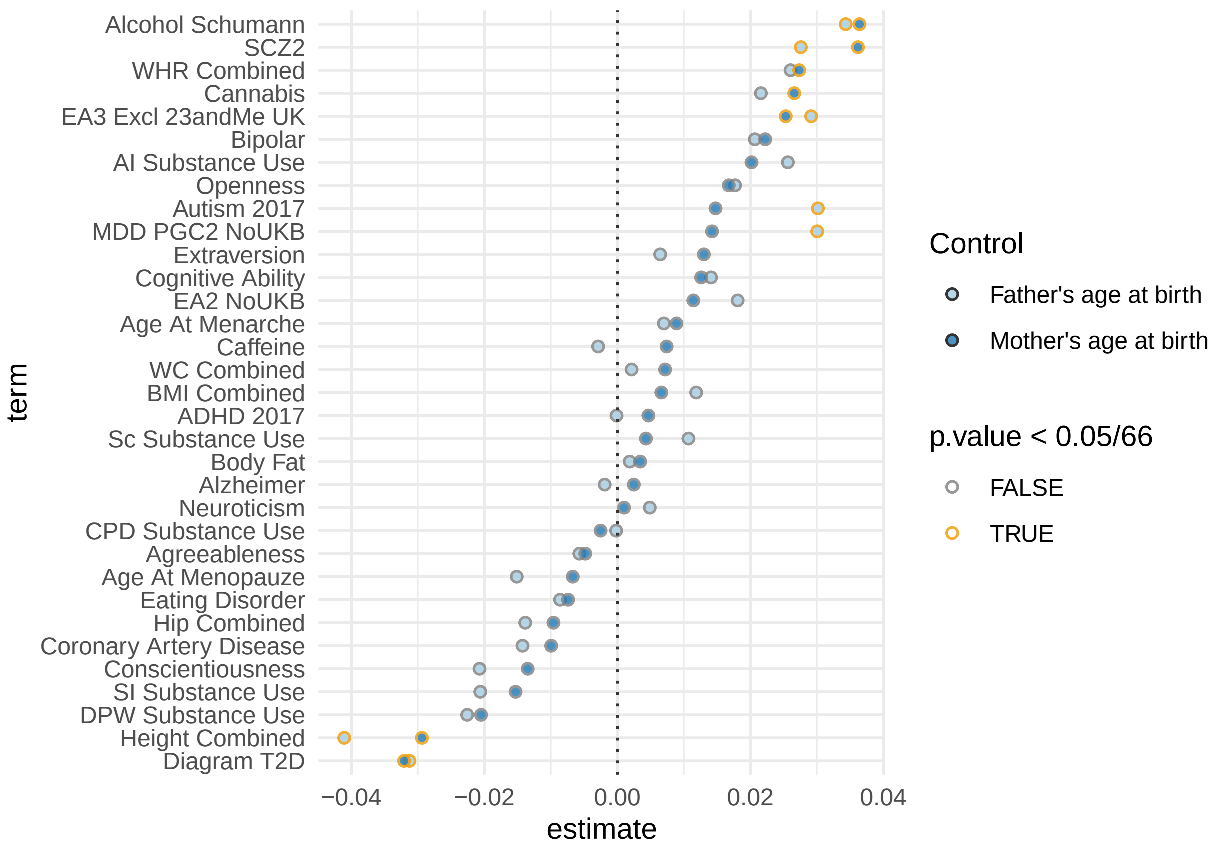


Figure 8: Effect sizes on number of children, controlling for age at first live birth (women only). Effect sizes for women without controls are shown for comparison



6.3 Number of sexual partners

Figure 9 splits males and females up by lifetime number of sexual partners. Remarkably, across both sexes, negative selection is strongly reversed for respondents who had 3 or fewer sexual partners in their lifetime.

6.4 Education and income

Figure 10 splits respondents up by education levels. Both negative and positive selection are typically larger and more significant for those who left school before 16. Table 4 summarizes the results.

Table 4: Negative selection by education level

Age left FTE	**% PGS significant**
< 16	63.6
16-18	27.3
> 18	6.06

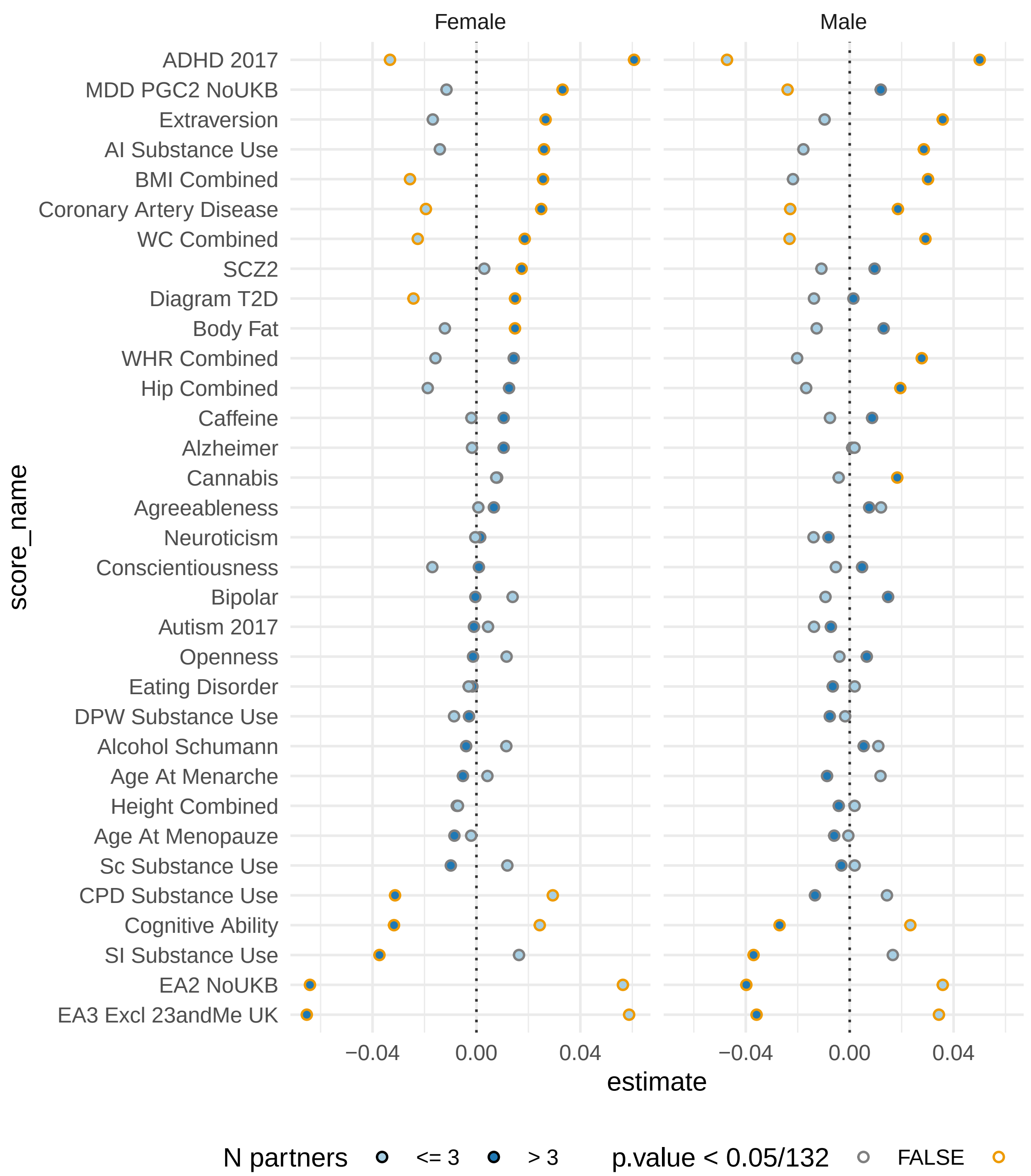


Figure 9: Effect sizes on number of children by number of sexual partners

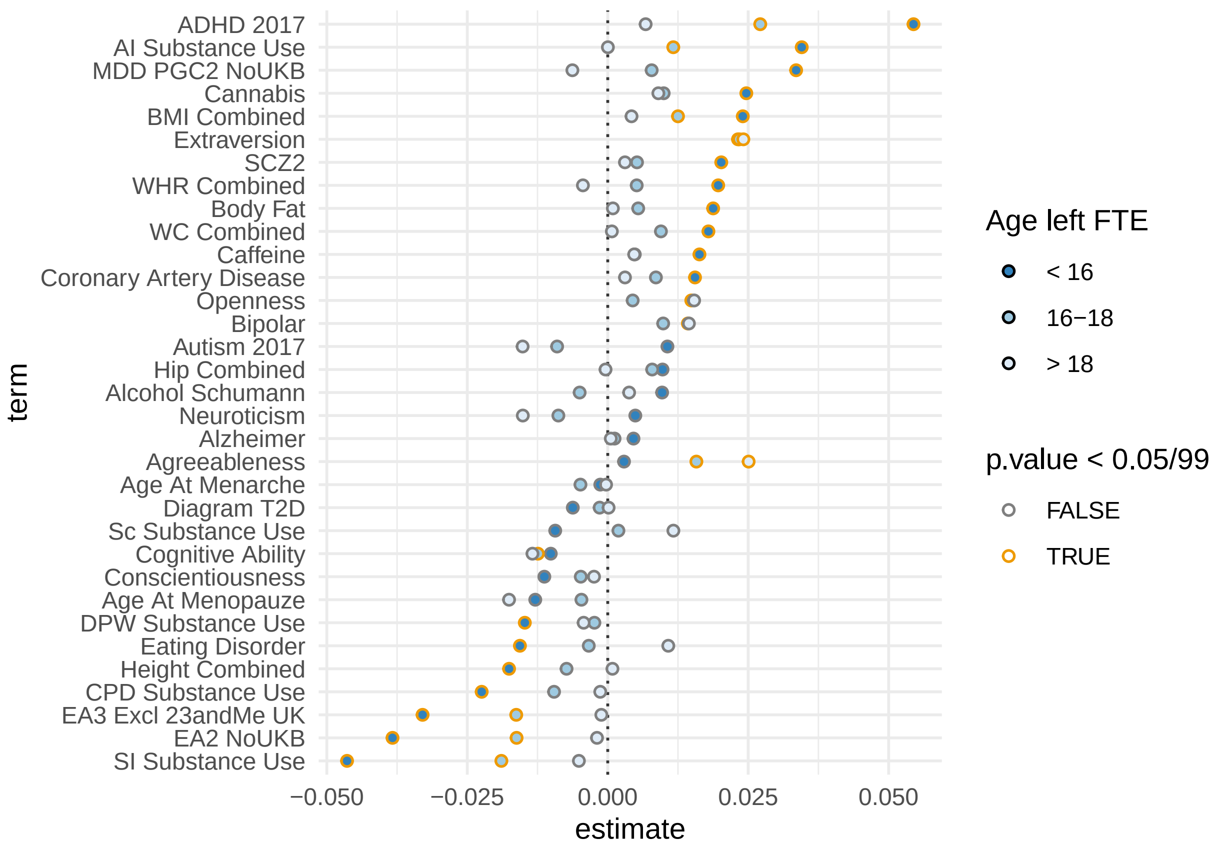


Figure 10: Effect sizes on number of children by age left full-time education

Figure 11 splits respondents by household income category. A very similar pattern holds, with selection effects being larger for those in the poorest income category.

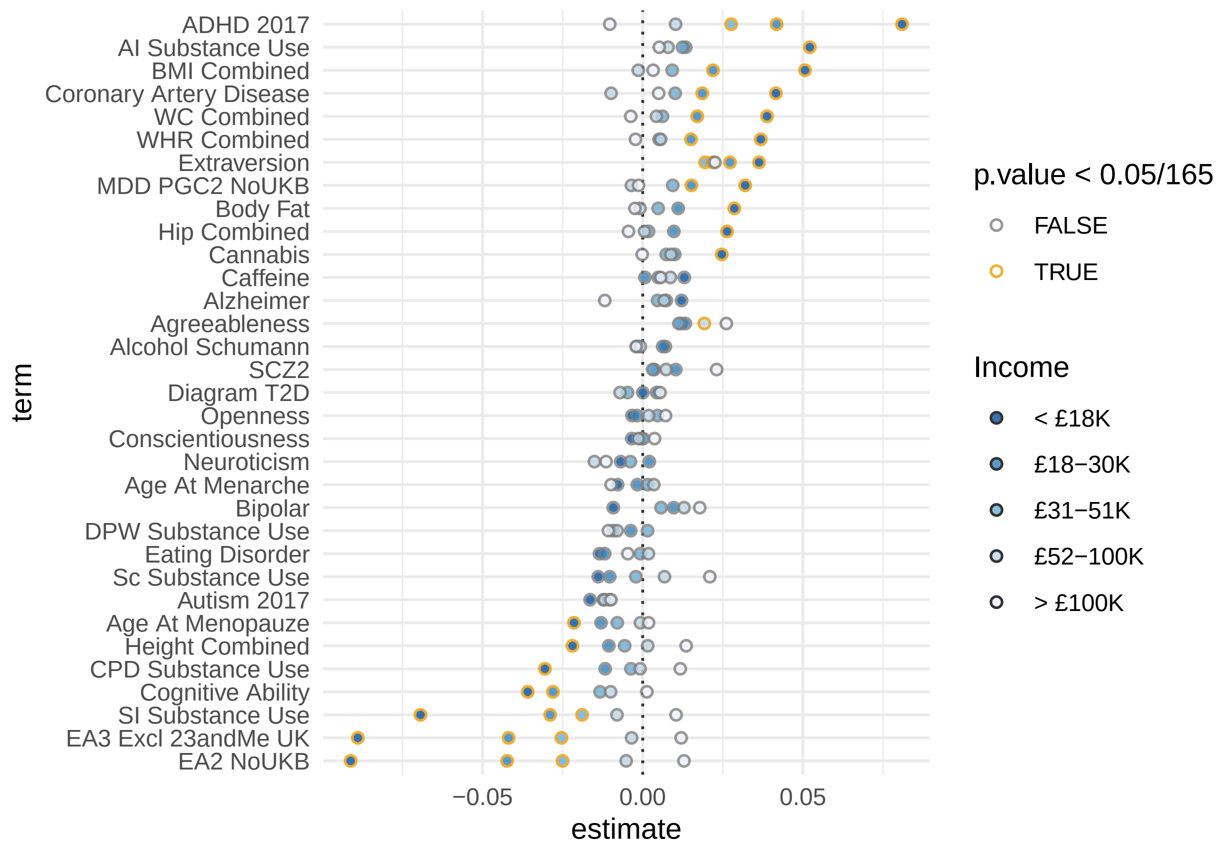


Figure 11: Effect sizes on number of children by household income

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.

7 Number of children

Figure 12 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.

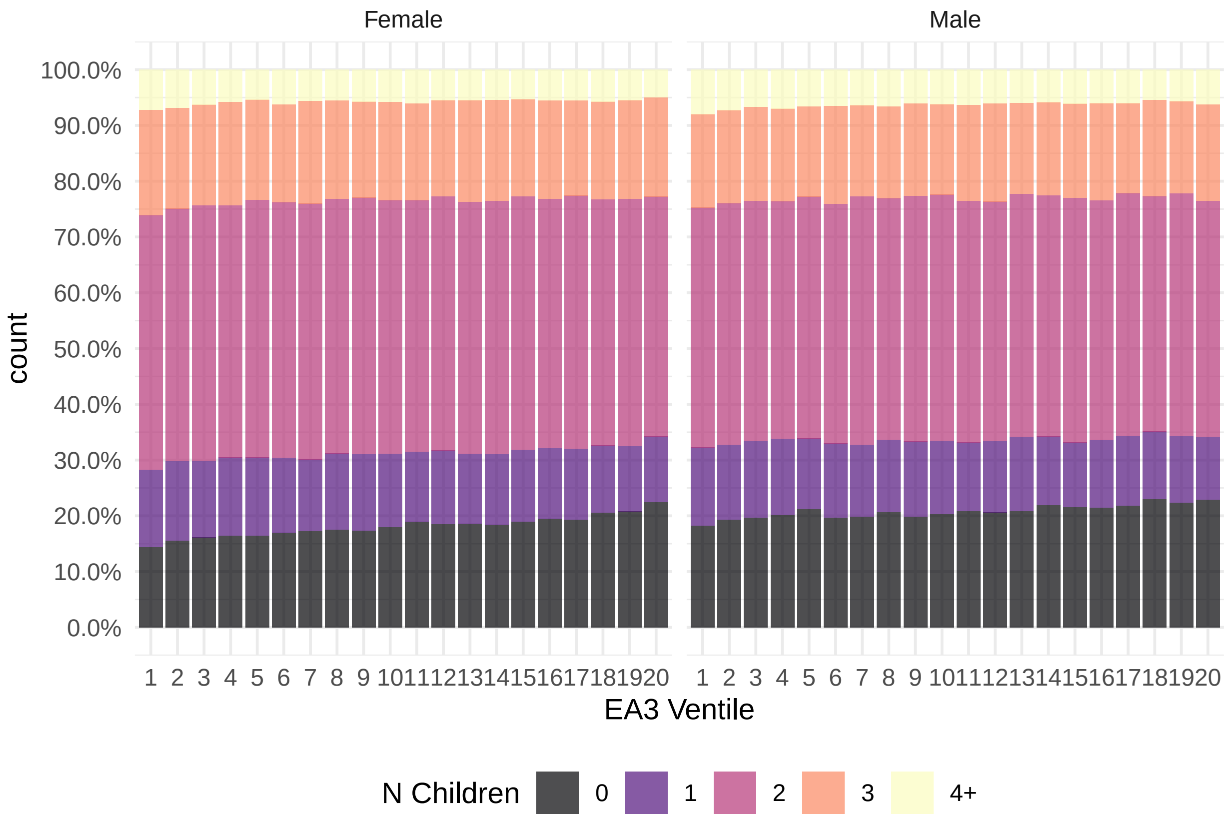


Figure 12: Number of children by ventiles of EA3 PGS