The Genetic Lottery for Premature Mortality in Mid-Century Wisconsin

Using the Phenotype Differences Model to Identify Genetic Effects with Incomplete Sibling Data

Sam Trejo

Assistant Professor
Department of Sociology
Office of Population Research

PRINCETON
UNIVERSITY

Klint Kanopka
Ph.D. Candidate, Stanford University

PRINCETON
UNIVERSITY

Introduction

- GWAS have mapped the genetic correlates of wide-range of complex traits
  - Results are used to generate PGI, which aim to index individual

- But do GWAS discoveries (and resulting PGI) capture the causal effects of genes?
  - Threat of environmental confounding from population stratification and dynastic effects
    - Young et al. 2019, Science
    - Okbay et al. 2022, Nature Genetics
    - Howe et al. 2022, Nature Genetics

Making Causal Inferences

- How do we identify causal genetic effects?

- Same as in non-genetic analyses
  - Leverage only random genetic variation

- With DNA, we have the ultimate “natural” experiment
  - Conditional on their parents’ genes, a child’s genes are randomly assigned via genetic recombination

Existing Methods

- Sibling methods
  - Family fixed effects difference out all shared family-level variation, indirectly conditioning on parental genotype
  - Requires siblings pairs with 2 genotypes & 2 phenotypes
- Trio methods
  - Explicitly conditions on parental genotype
  - Requires mother, father, & child trios with all 3 genotypes & the child's phenotype
    - Possible with only 2 genotypes child's phenotype using phased data

Limitations

- There is a dearth of the sort of genotyped family data required by FE and Trio Methods
  - UKB has 500k singletons but only has 16k sibling pairs & 10k parent-child pairs

Moving Forward

- How do we increase the sample sizes available for robust familial analyses?

- Introducing the Phenotype Differences Model!

- Requires only **one sibling’s genotype**, alongside two siblings’ phenotypes

Potential Applications

- Surveying individuals on the phenotypes of their siblings (e.g. the UKB is expanding)
- Merging phenotypic data of siblings from population registries, health records, etc.
- Using siblings pairs with missing data in existing longitudinal studies (e.g. in the WLS)

Potential Applications

- PD can both increase statistical power (by increasing sample size) and improve external validity (increasing representativeness of samples)

First Differences

$$y_{1j} - y_{2j} = \hat{\beta}^{\mathsf{FE}}(g_{1j} - g_{2j}) + \hat{\varepsilon}^{*}_{ij}$$

First Differences

$$y_{1j} - y_{2j} = \hat{\beta}^{\mathsf{FE}}(g_{1j} - g_{2j}) + \hat{\varepsilon}^{*}_{ij}$$

Phenotype Differences (General)

$$y_{1j} - y_{2j} = \hat{\alpha} + \hat{\beta}^{\mathsf{PD}}\left(g_{1j}(1 - \rho^{g_{1j},g_{2j}})\right) + \hat{\varepsilon}_{ij}$$

First Differences

$$y_{1j} - y_{2j} = \hat{\beta}^{\mathsf{FE}}(g_{1j} - g_{2j}) + \hat{\varepsilon}^{*}_{ij}$$

Phenotype Differences (General)

$$y_{1j} - y_{2j} = \hat{\alpha} + \hat{\beta}^{\mathsf{PD}}\left(g_{1j}(1 - \rho^{g_{1j},g_{2j}})\right) + \hat{\varepsilon}_{ij}$$

Phenotype Differences ($\rho^{g_{1j},g_{2j}} = .5$)

$$y_{1j} - y_{2j} = \hat{\alpha} + \hat{\beta}^{\mathsf{PD}}\frac{g_{1j}}{2} + \hat{\varepsilon}_{ij}$$

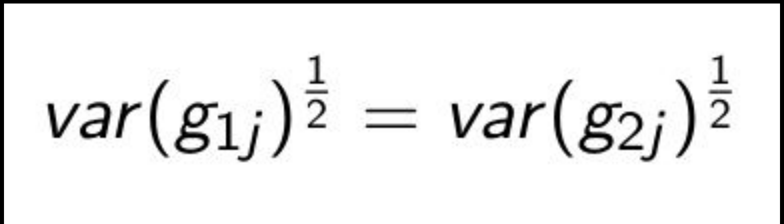
Comparative Efficiency

- When genetic effects are small (i.e. GWAS), Phenotype Differences provides the same precision as Fixed Effects *per genotype*
  - Though, you typically have half as many genotypes per family

- As genetic effects get larger, Phenotype Differences becomes comparatively less efficient than Fixed Effects per genotype
  - For current EA PGS, comparative precision drops from 1 to about 0.9

Key Assumption

- Equal genotype/PGI standard deviation of genetically observed and unobserved sibling

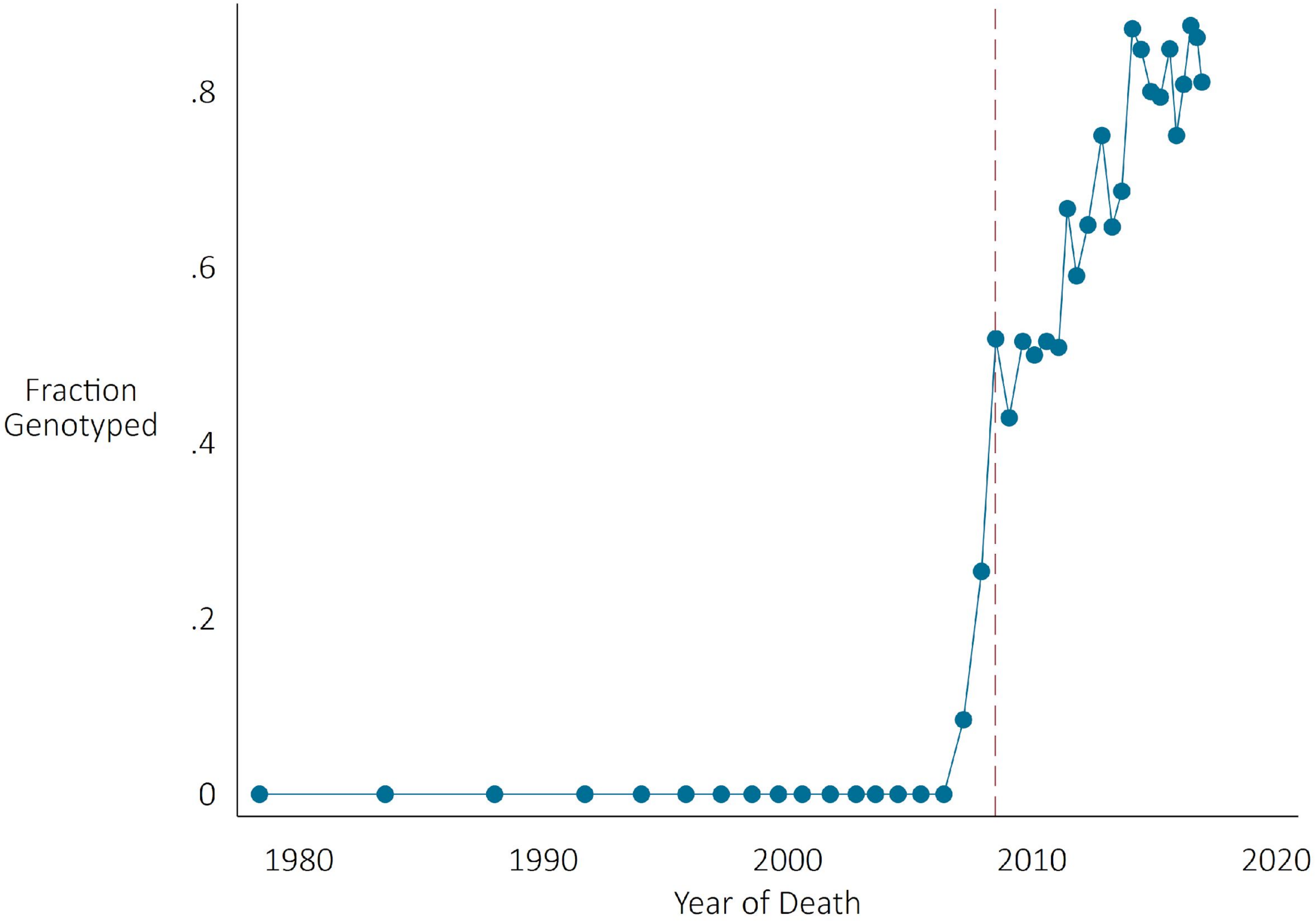


$$var(g_{1j})^{\frac{1}{2}} = var(g_{2j})^{\frac{1}{2}}$$

Not a problem for Phenotype Differences

- Asymmetric classical measurement error
  - E.g., respondents reporting their siblings' phenotype less accurately than their own
- Asymmetric measurement bias
  - E.g., respondents systematically under- or over-estimating their siblings' phenotype
- Linear selection into genotyping
  - E.g., genetic differences between individuals additively increasing or decreasingly likelihood of being the genotyped (versus ungenotyped) sibling

WISCONSIN
LONGITUDINAL
STUDY


Fraction Genotyped
0
.2
.4
.6
.8
1980
1990
2000
2010
2020
Year of Death

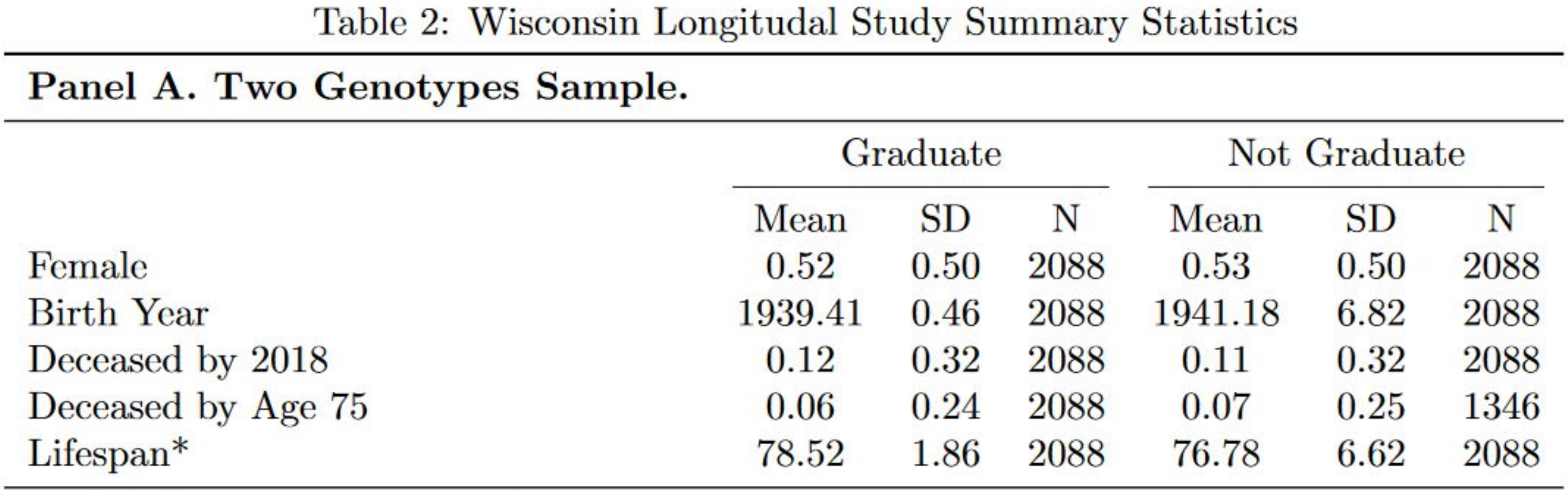


Table 2: Wisconsin Longitudal Study Summary Statistics

Panel A. Two Genotypes Sample.

	Graduate			Not Graduate		
	Mean	SD	N	Mean	SD	N
Female	0.52	0.50	2088	0.53	0.50	2088
Birth Year	1939.41	0.46	2088	1941.18	6.82	2088
Deceased by 2018	0.12	0.32	2088	0.11	0.32	2088
Deceased by Age 75	0.06	0.24	2088	0.07	0.25	1346
Lifespan*	78.52	1.86	2088	76.78	6.62	2088

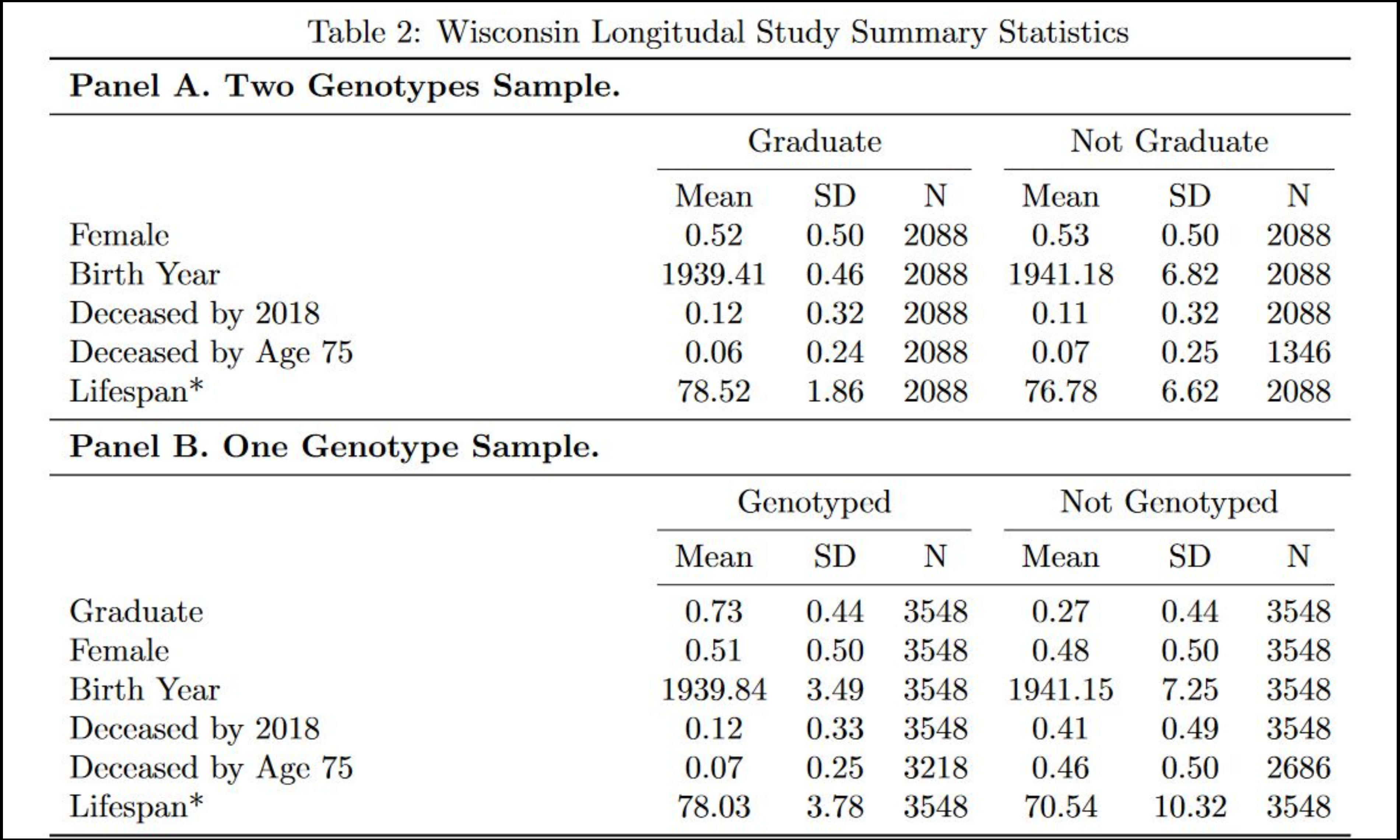


Table 2: Wisconsin Longitudal Study Summary Statistics

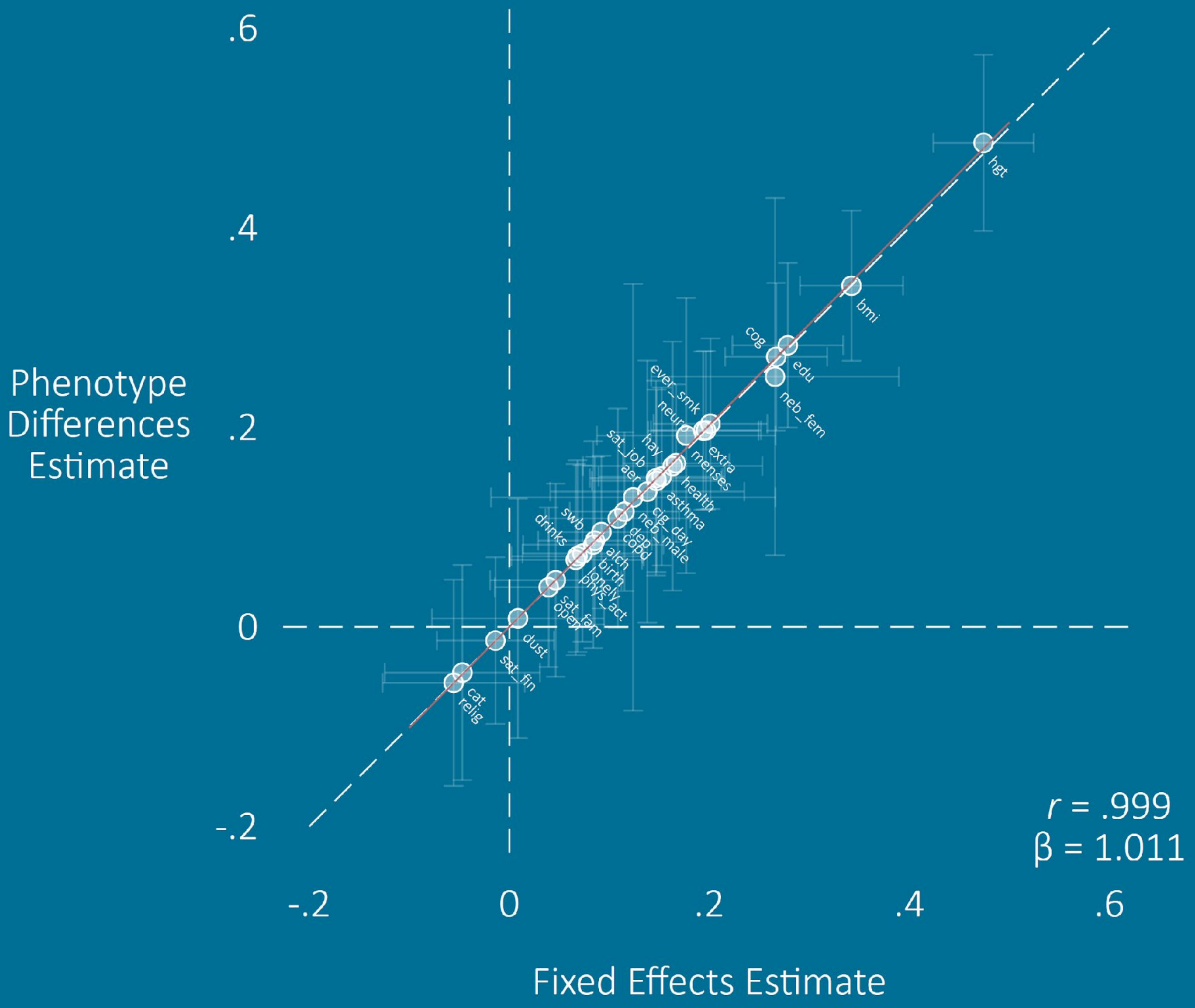
Panel A. Two Genotypes Sample.

	Graduate			Not Graduate		
	Mean	SD	N	Mean	SD	N
Female	0.52	0.50	2088	0.53	0.50	2088
Birth Year	1939.41	0.46	2088	1941.18	6.82	2088
Deceased by 2018	0.12	0.32	2088	0.11	0.32	2088
Deceased by Age 75	0.06	0.24	2088	0.07	0.25	1346
Lifespan*	78.52	1.86	2088	76.78	6.62	2088

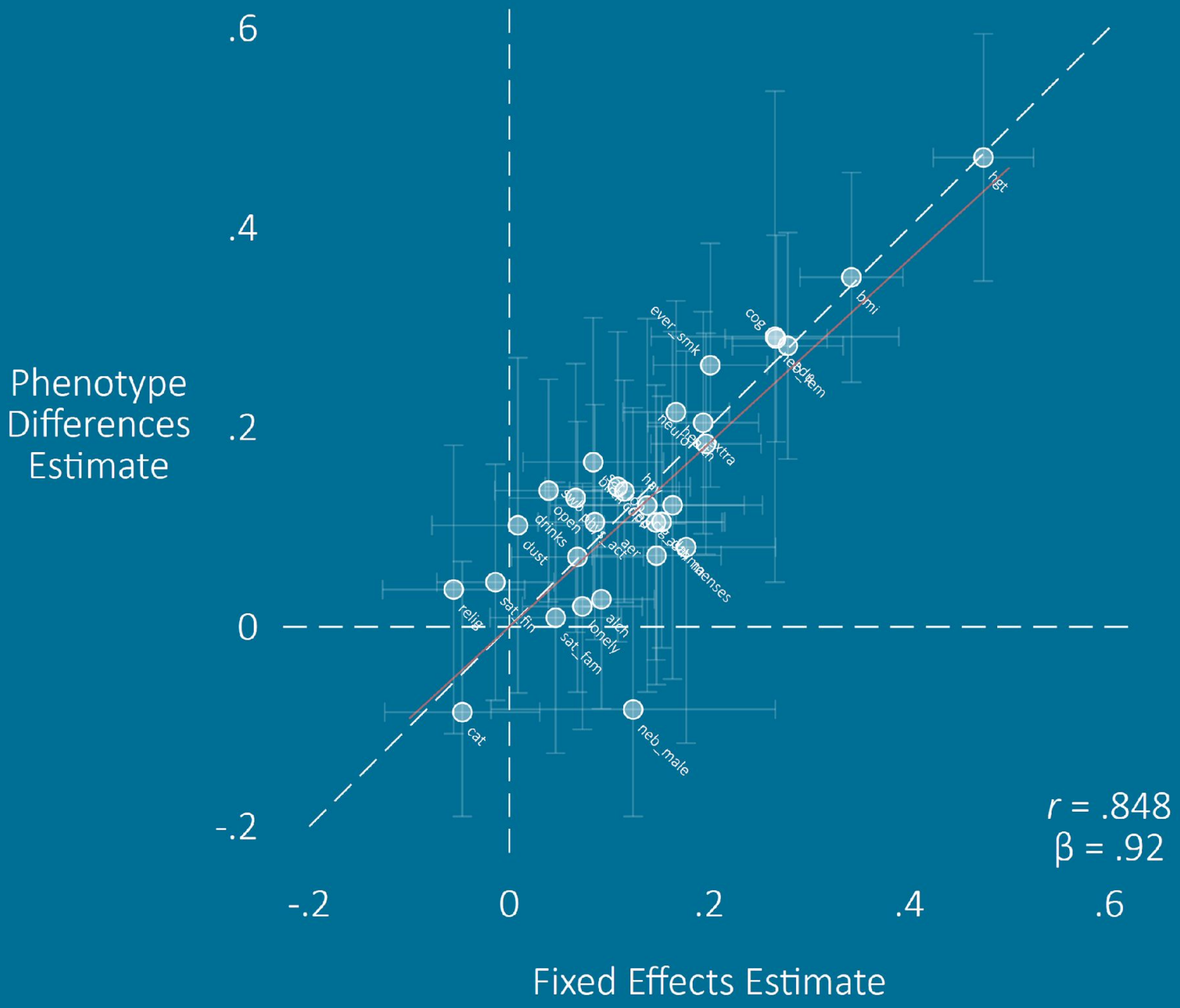
Panel B. One Genotype Sample.

	Genotyped			Not Genotyped		
	Mean	SD	N	Mean	SD	N
Graduate	0.73	0.44	3548	0.27	0.44	3548
Female	0.51	0.50	3548	0.48	0.50	3548
Birth Year	1939.84	3.49	3548	1941.15	7.25	3548
Deceased by 2018	0.12	0.33	3548	0.41	0.49	3548
Deceased by Age 75	0.07	0.25	3218	0.46	0.50	2686
Lifespan*	78.03	3.78	3548	70.54	10.32	3548

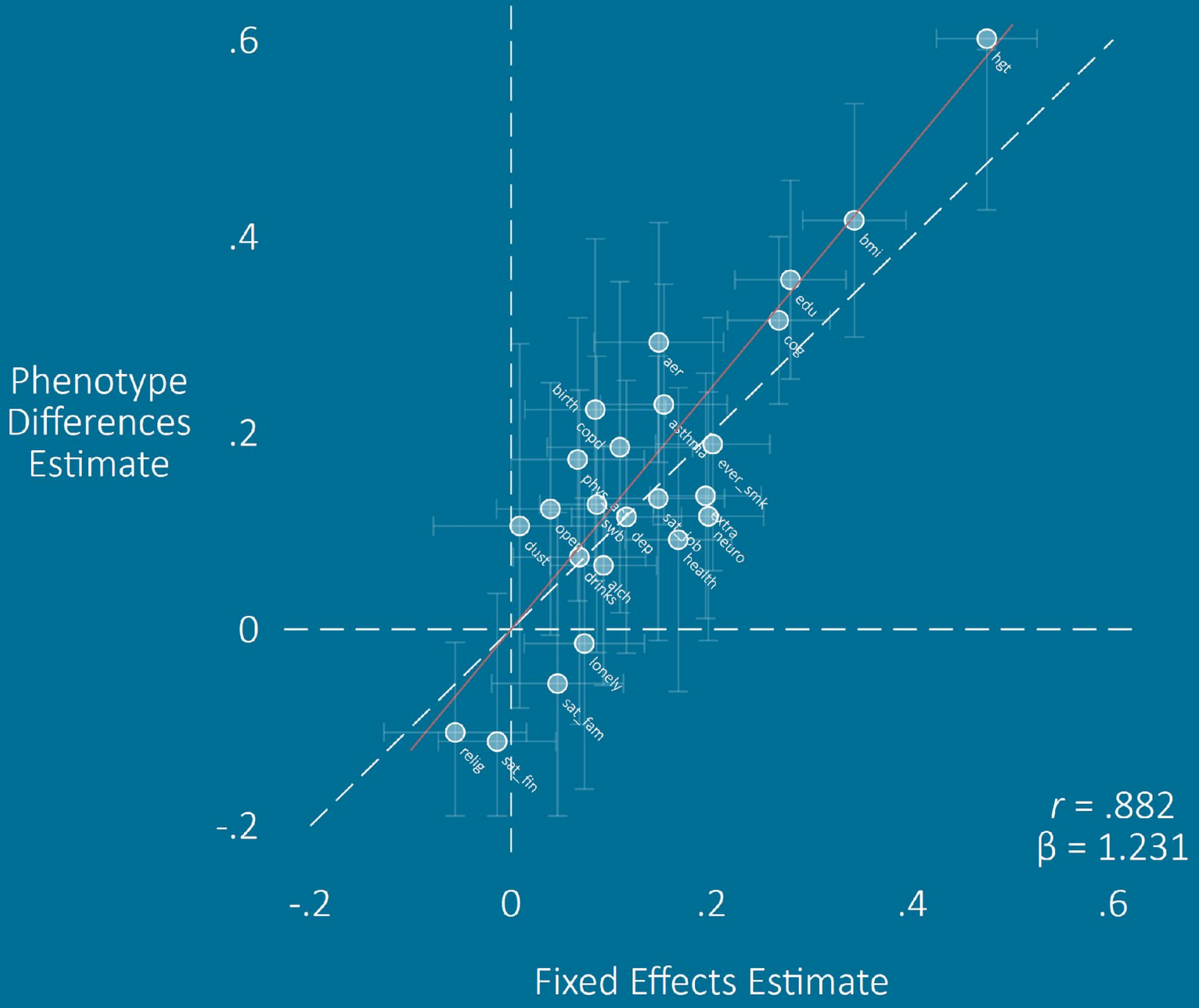
Full 2x Genotype Sample

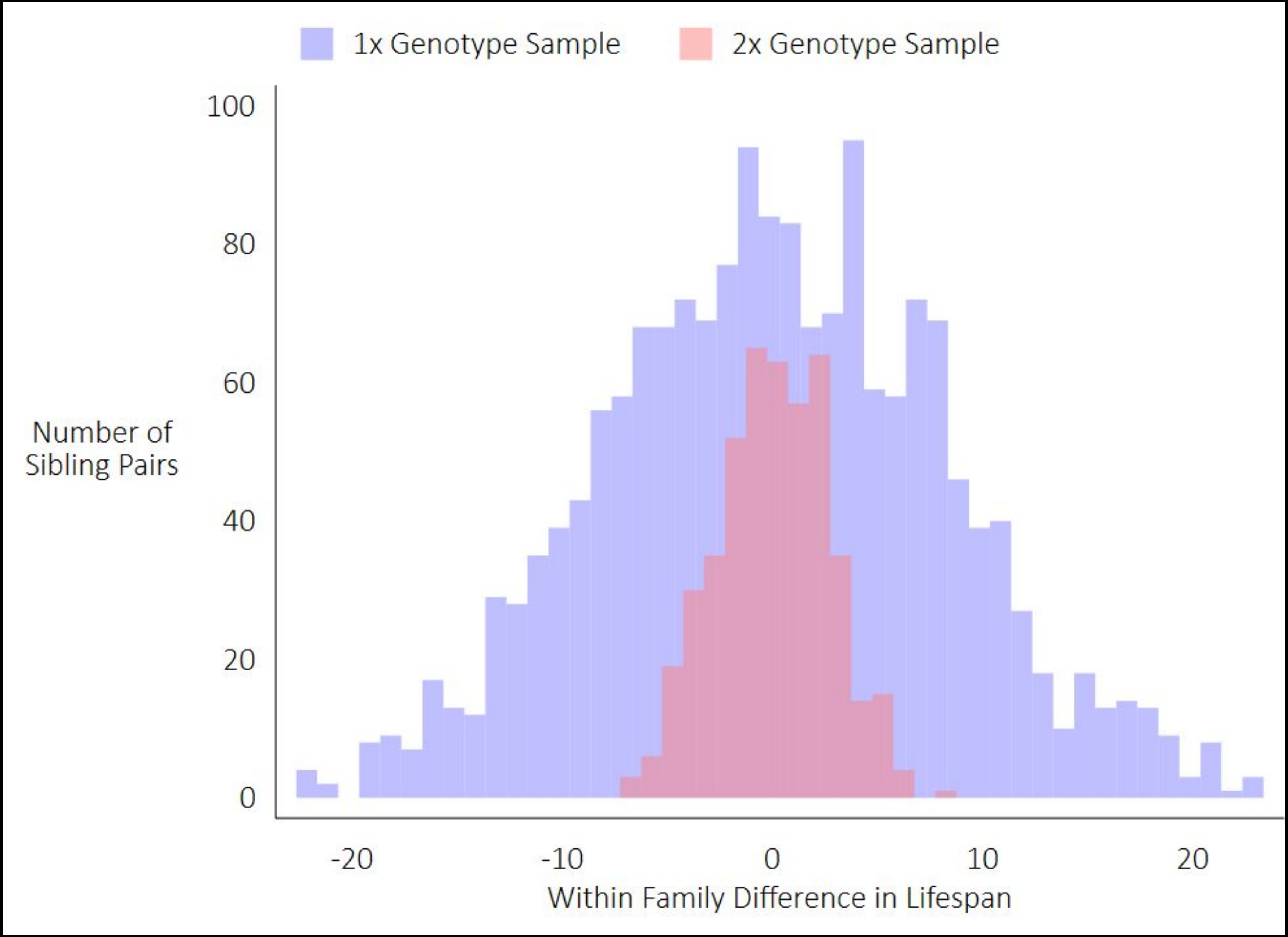


Half 2x Genotype Sample

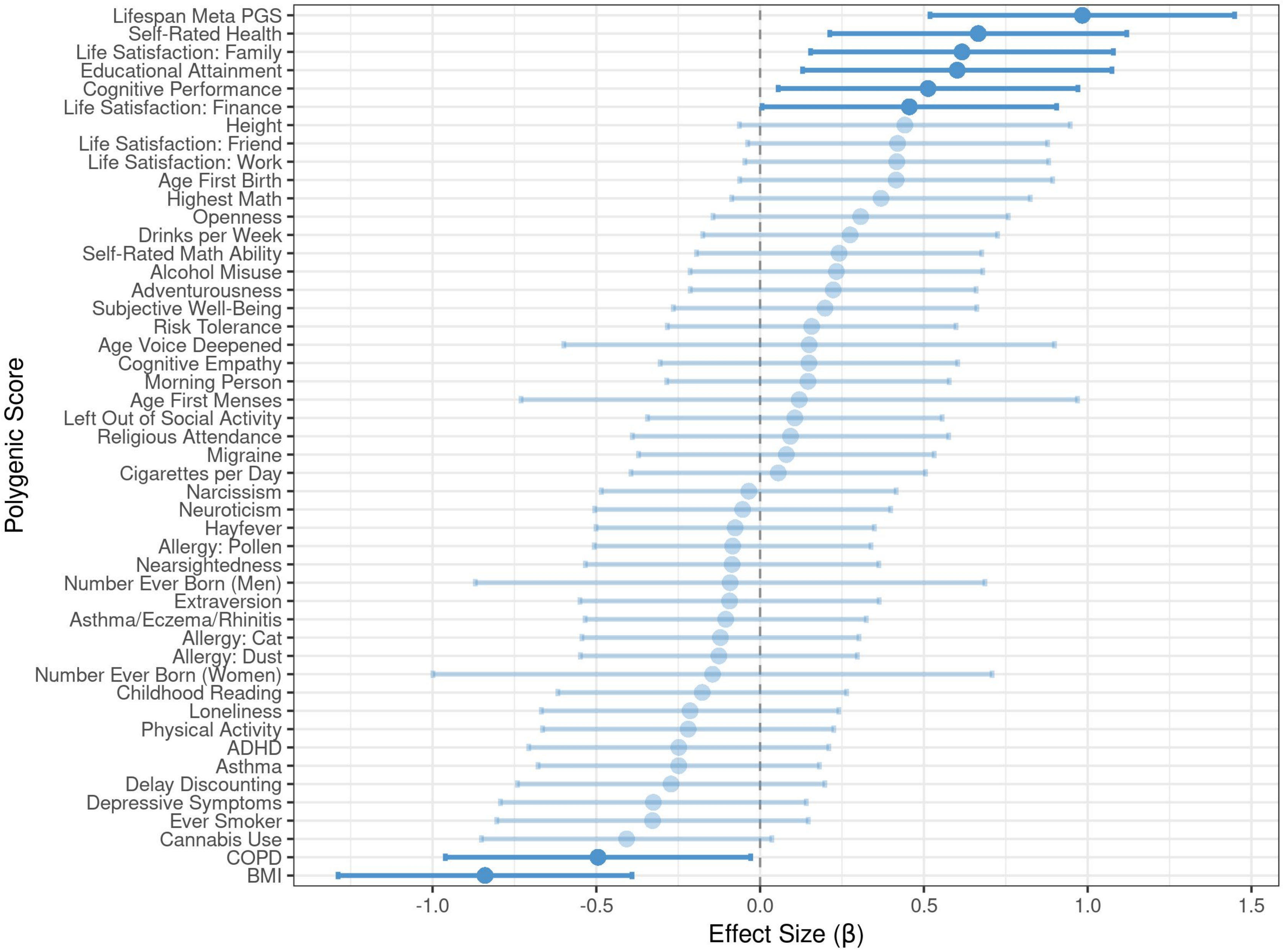


1x Genotype Sample

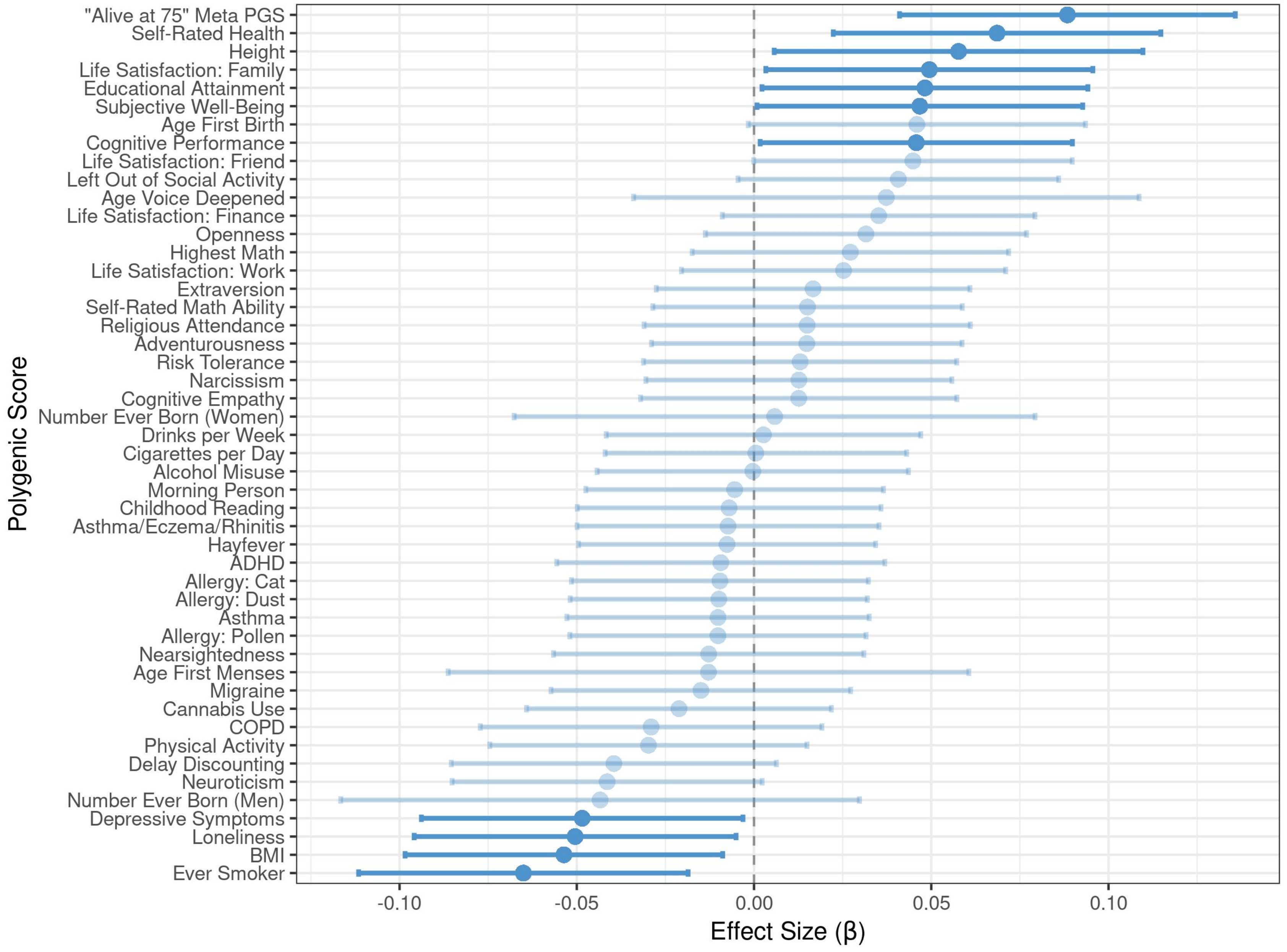




PRINCETON UNIVERSITY



PRINCETON
UNIVERSITY



PRINCETON
UNIVERSITY

Conclusion

- The Phenotype Differences model can increase power and external validity for the study of genetic effects
  - We need to collect more sibling phenotype data

- Twelve polygenic scores have statistically significant causal effects on mortality outcomes

Thanks!

www.samtrejo.com

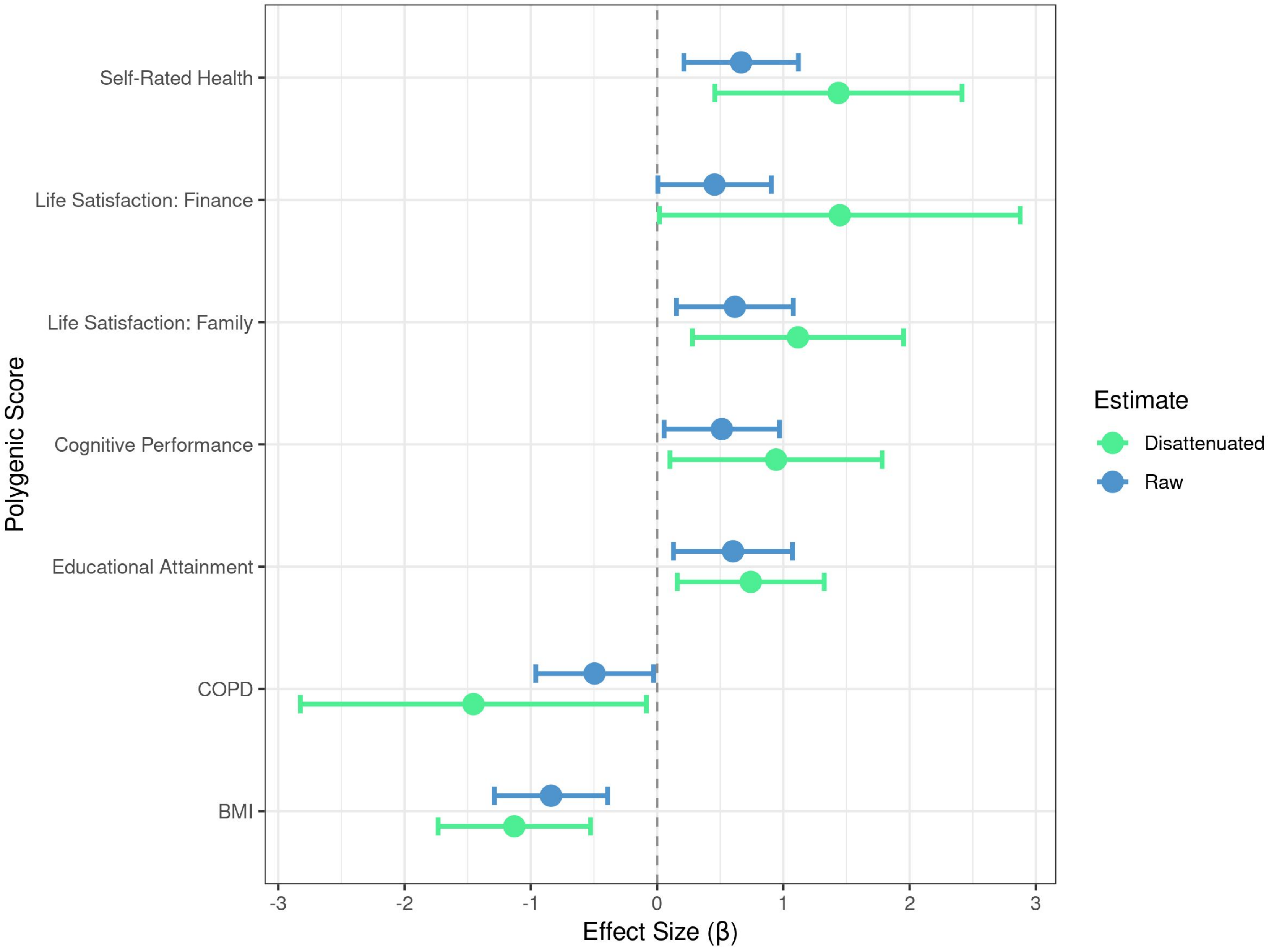
Fixed Effects

$$y_{ij} = \hat{\tau}_j + \hat{\beta}^{\mathsf{FE}} g_{ij} + \hat{\varepsilon}^*_{ij}$$

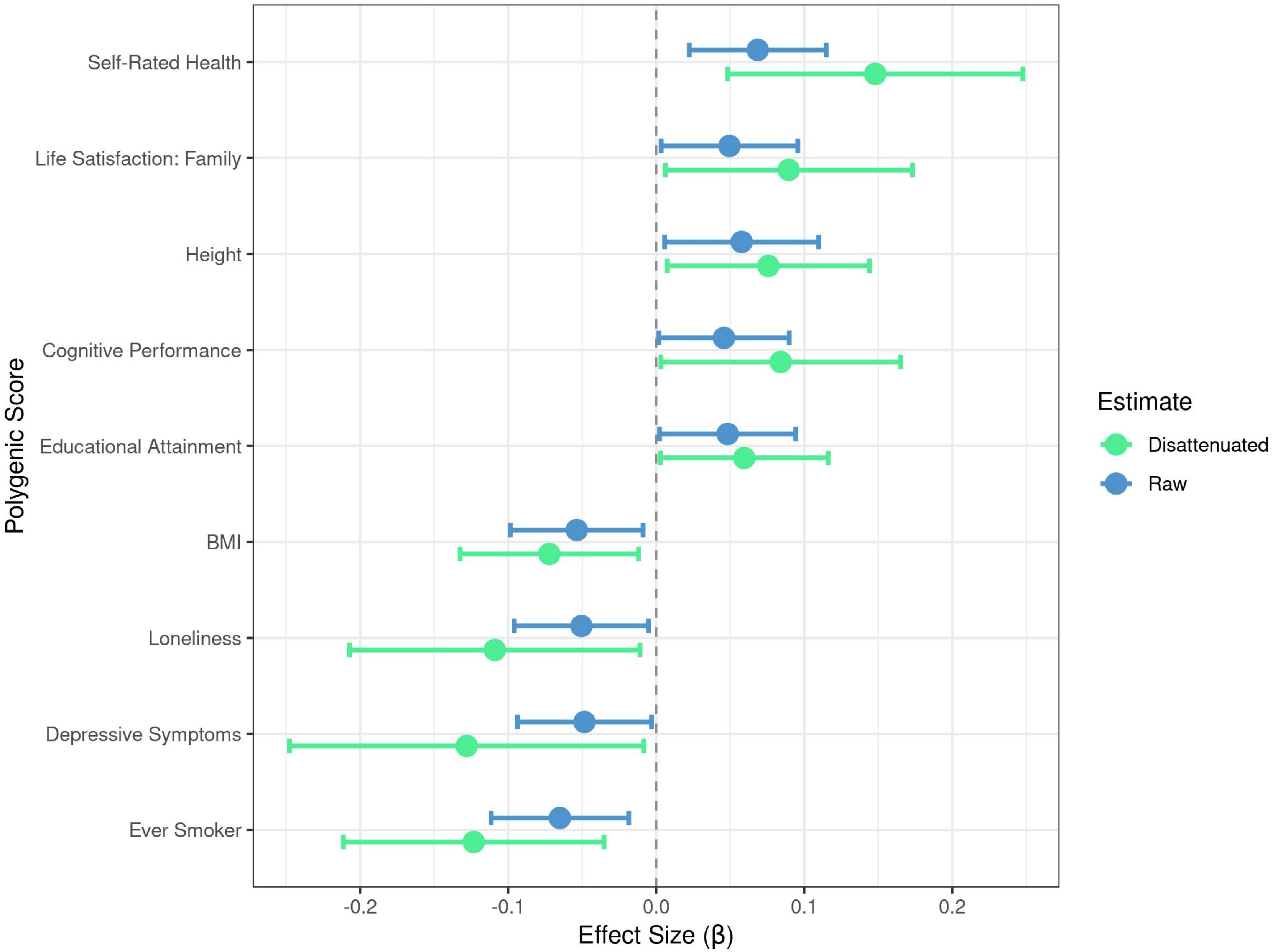
$$y_{ij} - \bar{y}_j = \hat{\beta}^{\mathsf{FE}}(g_{ij} - \bar{g}_j) + \hat{\varepsilon}^*_{ij}$$

First Differences

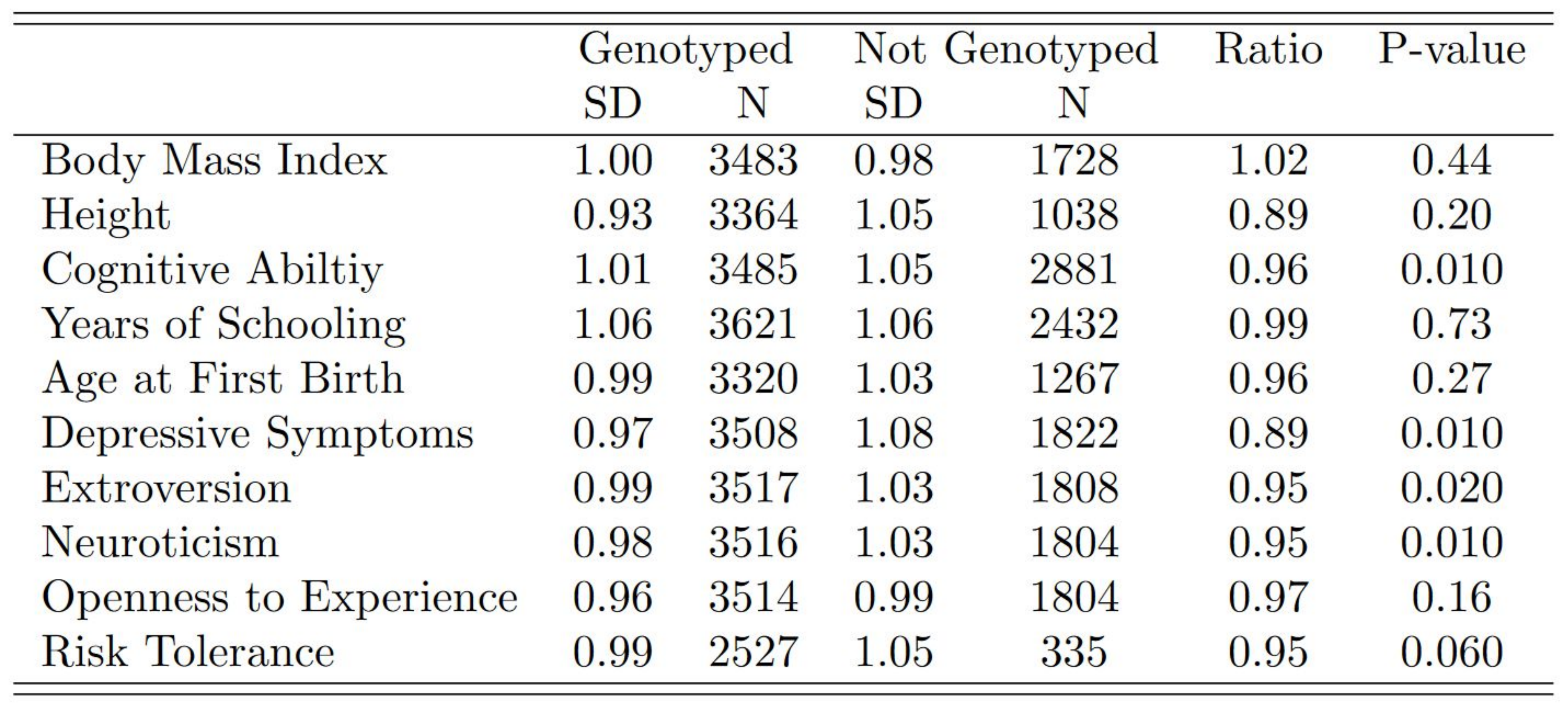
$$y_{1j} - y_{2j} = \hat{\beta}^{\mathsf{FE}}(g_{1j} - g_{2j}) + \hat{\varepsilon}^*_{ij}$$



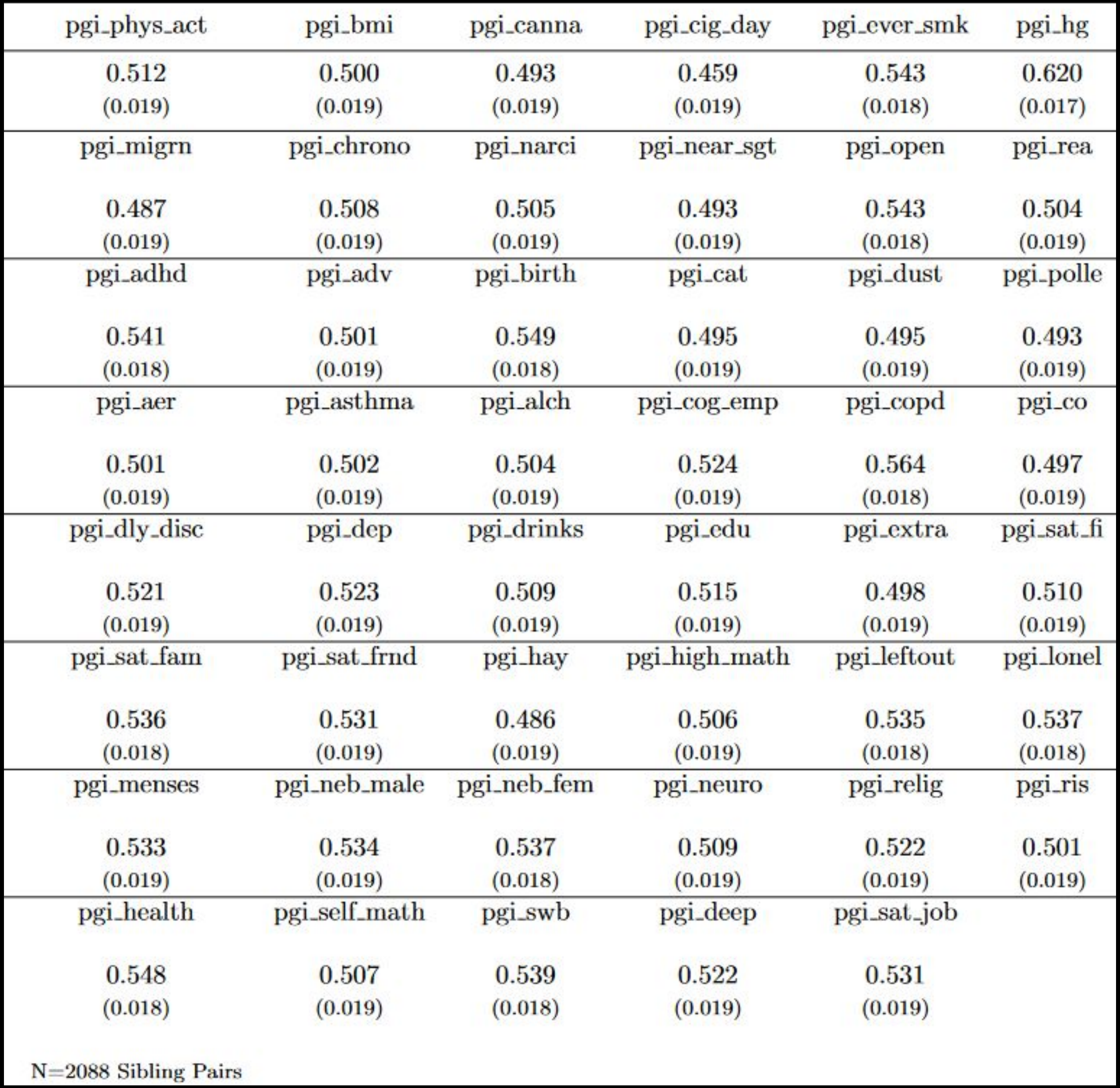
PRINCETON UNIVERSITY



PRINCETON UNIVERSITY

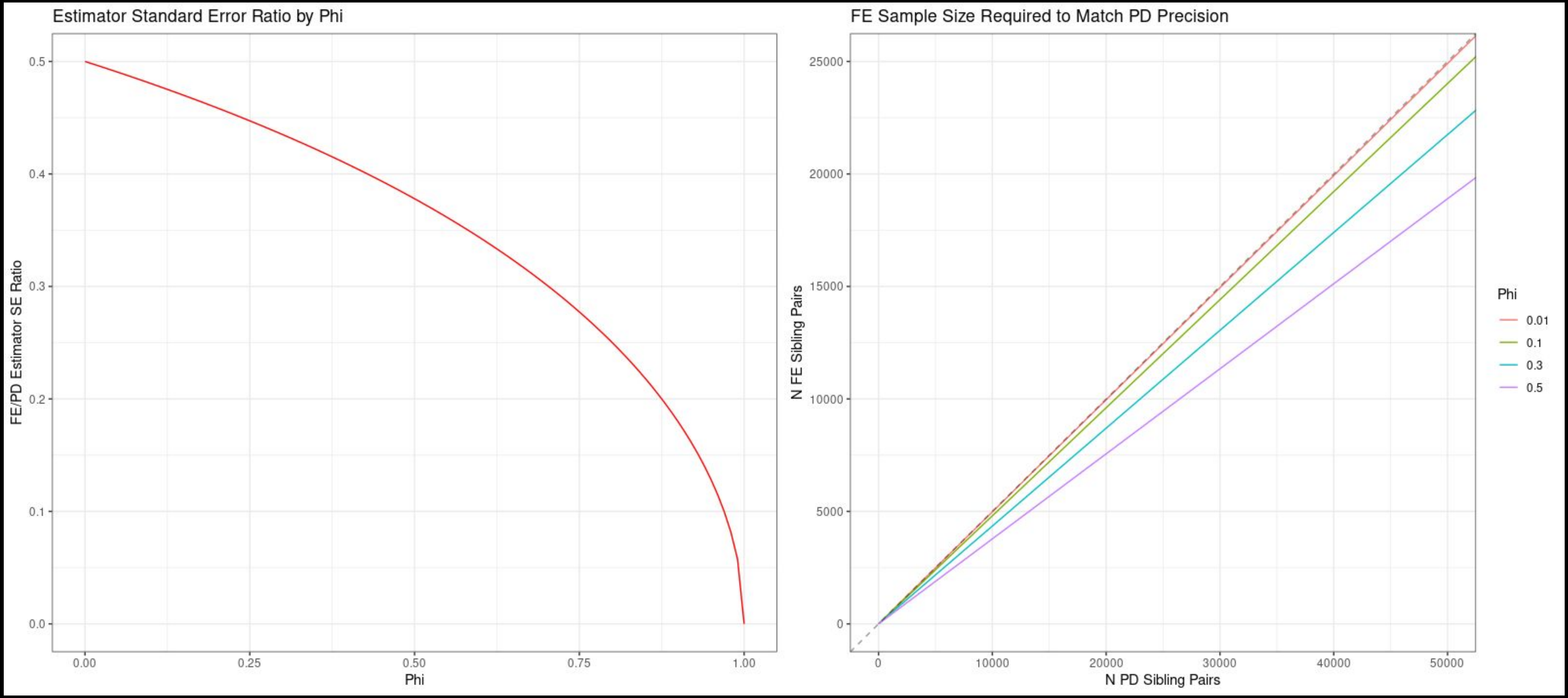


	Genotyped		Not Genotyped		Ratio	P-value
	SD	N	SD	N		
Body Mass Index	1.00	3483	0.98	1728	1.02	0.44
Height	0.93	3364	1.05	1038	0.89	0.20
Cognitive Abiltiy	1.01	3485	1.05	2881	0.96	0.010
Years of Schooling	1.06	3621	1.06	2432	0.99	0.73
Age at First Birth	0.99	3320	1.03	1267	0.96	0.27
Depressive Symptoms	0.97	3508	1.08	1822	0.89	0.010
Extroversion	0.99	3517	1.03	1808	0.95	0.020
Neuroticism	0.98	3516	1.03	1804	0.95	0.010
Openness to Experience	0.96	3514	0.99	1804	0.97	0.16
Risk Tolerance	0.99	2527	1.05	335	0.95	0.060



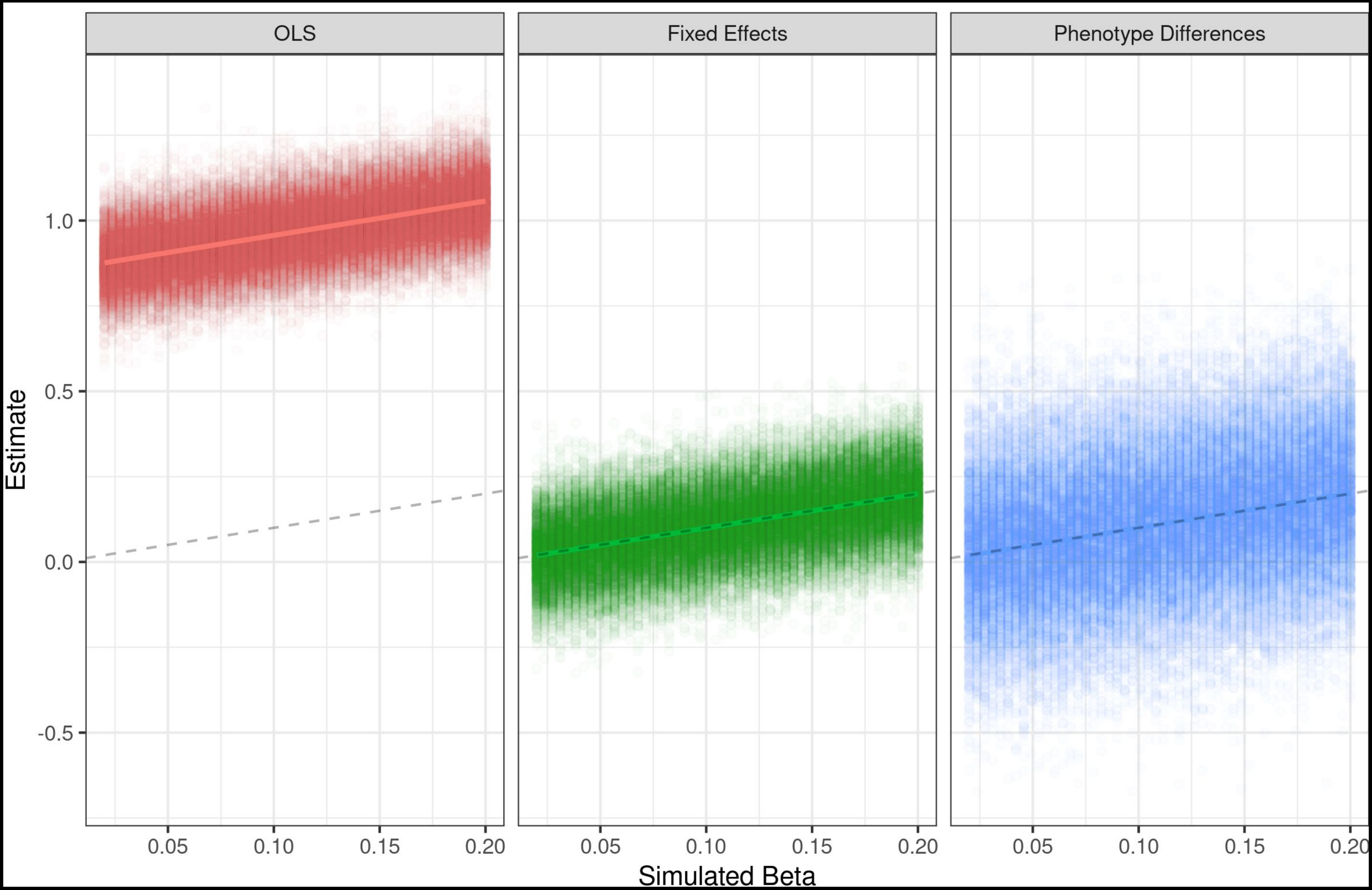
pgi_phys_act	pgi_bmi	pgi_canna	pgi_cig_day	pgi_ever_smk	pgi_hg
0.512 (0.019)	0.500 (0.019)	0.493 (0.019)	0.459 (0.019)	0.543 (0.018)	0.620 (0.017)
pgi_migrn	pgi_chrono	pgi_narci	pgi_near_sgt	pgi_open	pgi_rea
0.487 (0.019)	0.508 (0.019)	0.505 (0.019)	0.493 (0.019)	0.543 (0.018)	0.504 (0.019)
pgi_adhd	pgi_adv	pgi_birth	pgi_cat	pgi_dust	pgi_polle
0.541 (0.018)	0.501 (0.019)	0.549 (0.018)	0.495 (0.019)	0.495 (0.019)	0.493 (0.019)
pgi_aer	pgi_asthma	pgi_alch	pgi_cog_emp	pgi_copd	pgi_co
0.501 (0.019)	0.502 (0.019)	0.504 (0.019)	0.524 (0.019)	0.564 (0.018)	0.497 (0.019)
pgi_dly_disc	pgi_dep	pgi_drinks	pgi_edu	pgi_extra	pgi_sat_fi
0.521 (0.019)	0.523 (0.019)	0.509 (0.019)	0.515 (0.019)	0.498 (0.019)	0.510 (0.019)
pgi_sat_fam	pgi_sat_frnd	pgi_hay	pgi_high_math	pgi_leftout	pgi_lonel
0.536 (0.018)	0.531 (0.019)	0.486 (0.019)	0.506 (0.019)	0.535 (0.018)	0.537 (0.018)
pgi_menses	pgi_neb_male	pgi_neb_fem	pgi_neuro	pgi_relig	pgi_ris
0.533 (0.019)	0.534 (0.019)	0.537 (0.018)	0.509 (0.019)	0.522 (0.019)	0.501 (0.019)
pgi_health	pgi_self_math	pgi_swb	pgi_deep	pgi_sat_job	
0.548 (0.018)	0.507 (0.019)	0.539 (0.018)	0.522 (0.019)	0.531 (0.019)	

N=2088 Sibling Pairs



Estimator Standard Error Ratio by Phi

FE Sample Size Required to Match PD Precision



PRINCETON
UNIVERSITY