bioRxiv preprint doi: https://doi.org/10.1101/524850; this version posted July 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Genetic Nature or Genetic Nurture?
Quantifying Bias in Analyses Using Polygenic Scores

Sam Trejo[1*]
Benjamin W. Domingue[1]

June 2019

1. Graduate School of Education, Stanford University
* Send correspondence to samtrejo@stanford.edu.

Acknowledgements

We wish to thank Kathleen Mullan Harris and the Add Health staff for access to the restricted-use genetic data. We are grateful to Patrick Turley, Dalton Conley, attendees of the Stanford Genetics & Social Science journal club, seminar participants at the Integrating Genetics and Society 2018 Conference, and two anonymous referees for helpful comments. This work has been supported by the Russell Sage Foundation and the Ford Foundation under Grant No. 96-17-04, the National Science Foundation under Grant No. DGE-1656518, and by the Institute of Education Sciences under Grant No. R305B140009. Any opinions expressed are those of the authors alone and should not be construed as representing the opinions of any foundation.

bioRxiv preprint doi: https://doi.org/10.1101/524850; this version posted July 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Abstract

Summary statistics from a genome-wide association study (GWAS) can be used to generate a polygenic score (PGS). For complex, behavioral traits, the correlation between an individual's PGS and their phenotype may contain bias alongside the causal effect of the individual's genes (due to geographic, ancestral, and/or socioeconomic confounding). We formalize the recent introduction of a different source of bias in regression models using PGSs: the effects of parental genes on offspring outcomes, also known as genetic nurture. GWAS do not discriminate between the various pathways through which genes influence outcomes, meaning existing PGSs capture both direct genetic effects and genetic nurture effects. We construct a theoretical model for genetic effects and show that, unlike other sources of bias in PGSs, the presence of genetic nurture biases PGS coefficients from both naïve OLS (between-family) and family fixed effects (within-family) regressions. This bias is in opposite directions; while naïve OLS estimates are biased upwards, family fixed effects estimates are biased downwards. We quantify this bias for a given trait using two novel parameters that we identify and discuss: (1) the genetic correlation between the direct and nurture effects and (2) the ratio of the SNP heritabilities for the direct and nurture effects.

bioRxiv preprint doi: https://doi.org/10.1101/524850; this version posted July 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

1. Introduction

1a. Genomics & the Social Sciences

Spurred by the plummeting cost of DNA sequencing and technological developments in processing large amounts of genetic data, researchers have made great strides in connecting genes to biological and social outcomes in a replicable manner. The key tool is the genome-wide association study (GWAS); a GWAS uses genotype and phenotype data from many individuals to probe the relationship between a given trait and thousands of regions of the genome (Pearson and Manolio 2008). GWAS are conducted on a wide variety of outcomes, ranging from proximal, biological phenotypes, such as blood pressure (Giri et al. 2019) and height (Yengo, Sidorenko, et al. 2018), to distal, behavioral phenotypes, such as depression (Hyde et al. 2016; Okbay et al. 2016) and educational attainment (Lee et al. 2018).

Findings from GWAS are often used to generate a predictor—a polygenic score (PGS)—meant to summarize an individual's genetic predisposition for a given trait. PGSs offer great promise to social scientists interested in incorporating genes into biosocial models of human behavior (Belsky and Israel 2014). In the short term, PGSs may be used as control variables in studies of environmental effects (Rietveld et al. 2013), used in gene-environment interaction studies to probe whether genetic effects are environmentally contingent (Trejo et al. 2018; Barcellos, Carvalho, and Turley 2018), and used to better understand how genetic factors influence developmental processes (Belsky et al. 2016; Belsky et al. 2013). In the long run, PGSs might be used to identify those who would benefit most from early medical or educational interventions, i.e., for a developmental disorder like dyslexia (Martschenko, Trejo, and Domingue 2019).

1b. The Problem of Confounding

A point of emphasis is that the same technique, GWAS, is being used to map the genetic architecture of a diverse set of phenotypes. It is not obvious that the methodology used to identify the underlying genetics of proximal, biological phenotypes can be deployed without side effect to interrogate the genetics of complex, socially contextualized phenotypes. Especially in the case of traits like depression and educational attainment, it is critical that existing GWAS results be interpreted cautiously (Martschenko, Trejo, and Domingue 2019); while PGSs have been shown to predict complex phenotypes, the relationship between an individual's PGS captures a broad range of information and associations with downstream outcomes and therefore cannot be readily interpreted as the causal effect of genes. An individual's genome contains fine-grain information about their place in the intricate structure of a population (Hamer and Sirota 2000; Novembre et al. 2008), meaning that GWASs for complex traits may simply identify genes related to confounding environmental variables such as ancestry, geography, or socioeconomic status.

Recent work in human studies has begun to elucidate a novel source of confounding: social genetic effects (Domingue and Belsky 2017). Social genetic effects, also known as indirect genetic effects, are defined as the influence of one organism's genotype on a different

bioRxiv preprint doi: https://doi.org/10.1101/524850; this version posted July 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

organism's phenotype. The idea of social genetic effects originated in evolutionary theory (Moore, Brodie, and Wolf 1997; Wolf et al. 1998), and social genetic effects have been observed in animal populations (Petfield et al. 2005; Bergsma et al. 2008; Canario, Lundeheim, and Bijma 2017; Baud et al. 2018). Social science is now beginning to study such effects in human populations; examples include among social peers (Sotoudeh, Conley, and Harris 2017; Domingue et al. 2018), sibling pairs (Cawley et al. 2017; Kong et al. 2018), and parents and their children (Bates et al. 2018; Kong et al. 2018; Wertz et al. 2018). The existence of within-family social genetic effects complicates attempts to derive causal estimates from GWAS.

For recent breakthroughs in the genetic architecture of complex traits to provide novel value to researchers in the biomedical and social sciences, the relationships discovered in a GWAS must mostly reflect causal relationships between an individual's genes and their phenotype. If, for example, the genes identified for a complex trait predict it only through spurious correlation, PGSs will provide little use towards broadening our understanding of genetic and environmental influences. Thus, validating PGSs within-families is vitally important for sifting out causation from correlation among the genetics identified in GWASs of complex traits (Rietveld et al. 2014; Domingue et al. 2015; Lee et al. 2018; Belsky et al. 2018). Environmental differences are muted between siblings and, conditional on parental genotype, child genotype is randomly assigned through a process known as genetic recombination (Conley and Fletcher 2017). This makes family fixed effect regression models that compare genetic differences in siblings to phenotypic differences in siblings the gold standard for testing and understanding whether genetic differences are causally related to downstream outcomes. Within-family research designs, however, are not without their own complications. Genetic nurture may lead to bias in estimates derived from within-family studies, though the extent of this bias has not yet been explored.

1c. Accounting for Genetic Nurture

In this paper, we describe how genetic nurture influences PGS construction and leads to bias within-family and between-family regression analyses using PGSs. We construct a theoretical model for additive genetic effects and show that, unlike other sources of bias in PGSs, the presence of genetic nurture can bias PGS coefficients from both naïve OLS (between-family) regressions and family fixed effects (within-family) regressions. We quantify the magnitude of this bias for a given trait using two novel parameters: (1) the genetic correlation between the direct and nurture effects and (2) the ratio of the SNP heritabilities for the direct and nurture effects. Bias is in opposite directions; whereas naïve OLS estimates are biased upwards, family fixed effects estimates are biased downwards. These findings highlight a shortcoming of existing PGSs and have important implications for the use and interpretation of research designs using PGSs for traits where genetic nurture is a relevant causal pathway.

The paper will proceed as follows. In Section 2, we motivate our theoretical model using empirical data. We introduce our theoretical model in Section 3 and then demonstrate its

bioRxiv preprint doi: https://doi.org/10.1101/524850; this version posted July 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

implications for GWAS and regression models using PGSs in Section 4. In Section 5, we discuss the insights gleaned from our theoretical model.

2. Empirical Motivation

2a. Empirical Model

We motivate our theoretical framework by first considering the empirical specifications used in recent work (Domingue et al. 2015; Belsky et al. 2018; Lee et al. 2018). Consider the following two models relating an individual's PGS constructed from recent GWAS results ($\widehat{PGS}'^{\mathrm{D}}_{ij}$) to their outcome ($Y_{ij}$):

$$\text{Model 1: } Y_{ij} = \hat{\psi}_0 + \hat{\psi}_1 \widehat{PGS}'^{\mathrm{D}}_{ij} + \overrightarrow{X_{ij}}\overrightarrow{\widehat{\Theta}} + \epsilon_{ij}$$

$$\text{Model 2: } Y_{ij} = \hat{\pi}_0 + \hat{\pi}_1 \widehat{PGS}'^{\mathrm{D}}_{ij} + \overrightarrow{X_{ij}}\overrightarrow{\widehat{\Theta}} + \Gamma_j + \epsilon_{ij}$$

(2a.i)

$\widehat{PGS}'^{\mathrm{D}}_{ij}$: Normalized PGS constructed from the observed linear relationship between genotype and outcome

Y_{ij}: Outcome for individual i in family j

Γ_j: Family j fixed effect

$\overrightarrow{X_{ij}}$: Vector of individual covariates $\left(\overrightarrow{\widehat{\Theta}}\right)$ comprised of sex, age, and the first 10 principal components of genotype

Model 1 treats individuals as though they are unrelated whereas Model 2 compares siblings using a family fixed effect. In effect, Model 2 asks whether sibling differences in PGS translate into sibling differences in the outcome. Thus, Model 1 leverages covariation in Y_{ij} and $\widehat{PGS}'^{\mathrm{D}}_{ij}$ between individuals from different families while Model 2 leverages only covariation in Y_{ij} and $\widehat{PGS}'^{\mathrm{D}}_{ij}$ between pairs of individuals in the same family.

2b. Unresolved Questions

Table 1 displays results from Model 1 and Model 2 using data from the National Longitudinal Study of Adolescent to Adult Health (Harris 2013) for six phenotypes: educational attainment (Lee et al. 2018), cognitive ability (Lee et al. 2018), depressive symptoms (Turley et al. 2018), birth weight (Warrington et al. 2019), body mass index (Locke et al. 2015), and height (Wood et al. 2014). We further discuss the Add Health data and PGS construction (including links to the GWAS summary statistics used) in Sections A1 and A2 of the appendix.

[Insert Table 1 Here]

In Table 1, a one standard deviation increase in the educational attainment PGS is associated with over an additional .8 year of schooling between-families ($\hat{\psi}_1$) but less than half of that within-families ($\hat{\pi}_1$). If we compare the six phenotypes, the relative size of $\hat{\psi}_1$ and $\hat{\pi}_1$ (captured by $\frac{\hat{\pi}_1}{\hat{\psi}_1}$) varies dramatically. For years of schooling and cognitive performance,

bioRxiv preprint doi: https://doi.org/10.1101/524850; this version posted July 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

bootstrapped p-values show that the differences seen within and between-family are statistically significant (i.e. $\hat{\psi}_1 \neq \hat{\pi}_1$). Moreover, these findings have been replicated by other researchers; a recent study using data from the United Kingdom found that PGS coefficients for cognitive traits were on average 60% greater between families than within-families (Selzam et al. 2019). They, like us, found no statistically significant differences between coefficients derived from within and between-family models for non-cognitive traits.

Why might this be the case? One possibility is that the between-family models are confounded while the within-family models capture the true causal effects of the PGS. Alternatively, it may be that some of the processes captured by GWAS function differently within-families versus between-families (for example, genetic nurturance). Answering this question requires a more formal treatment in order to better understand what underlying features may drive differences between $\hat{\psi}_1$ and $\hat{\pi}_1$ across phenotypes. Our theoretical model, which we develop below, suggests that, in addition to possible environmental confounding, bias in $\hat{\psi}_1$ and $\hat{\pi}_1$ can depend on two novel parameters: (1) the underlying genetic correlation of direct and nurture effects and (2) the ratio of the SNP heritabilities for direct effects and nurture effects.

3. Theoretical Model

3a. Direct Genetic Effects and Genetic Nurture Effects

Historically, biosocial analyses have modeled complex traits as a function of both direct genetic effects and environment influences on an individual. Motivated by recent work highlighting the relevance of genetic nurture effects (Bates et al. 2018; Kong et al. 2018; Belsky et al. 2018; Wertz et al. 2018), we extend this model to include the genes of an individual's parent. Thus, we assume outcome Y_{ij} is a function of individual i's genotype, the genotypes of the parents in family j, and distinct individual-level and family-level environments. We choose to have a common effect of parental genetics at a given loci, instead of separate maternal and paternal effects, given the lack of strong empirical evidence of differences across parents (Kong et al. 2018).

$$Y_{ij} = \beta_0 + f(G_{ij}) + f(G_j) + f(E_{ij}) + f(E_j) + \epsilon_{ij}$$

(3a.i)

$f(G_{ij})$: Effect of i's genome on Y_{ij}

$f(G_j)$: Effect of family j's genome on Y_{ij}

$f(E_{ij})$: Effect of i's environment on Y_{ij}

$f(E_j)$: Effect of family j's environment on Y_{ij}

Note that here the environmental components E_{ij} and E_j are defined as the strictly non-genetic sources of variation in Y_i. In other words, an environment influences a child's Y_i irrespective of their or their parents' genetic composition. Environmental features of a child's environment, such as their family's socioeconomic status, are captured in $f(G_j)$ if those environmental features were caused by their parents' genes. We make three important

bioRxiv preprint doi: https://doi.org/10.1101/524850; this version posted July 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

assumptions to simplify the exposition of this model: no gene-environment interaction, no gene-environment correlation, and no assortative mating (see Section A3 of the appendix for a mathematical expression of these assumptions). We discuss the likely implications of a violation of these assumptions for our results at the end of Section 5a. While some of these assumptions are unlikely to hold true in the real world, we emphasize that goal of our theoretical model is not to create a model that perfectly describes reality but rather to formalize the ways in which social genetic effects influence GWASs, PGSs, and inevitably the interpretation of findings in the field of social science genomics. Our model, while simple, illustrates the key empirical phenomenon of interest; these higher-order features of the real world should not change the key implications derived from our model.

3b. True Polygenic Scores

Complex, behavioral traits are associated with many genes across the genome that simultaneously produce very small effects (Chabris et al. 2015; Visscher et al. 2017). To increase statistical power and simplify computation, researchers often summarize the relevant genetics of individual i into a single linear predictor called a PGS (Dudbridge 2013). This has become a widely utilized technique (Duncan et al. 2018) and relies on the assumptions that genetic effects are linear and additive. Recent meta-analyses of twin studies support the linear, additive model for genetic effects (Polderman et al. 2015). For the remainder of the paper, we approximate $f\left(G_{ij}\right)$ and $f\left(G_j\right)$ in our theoretical model (3a.i) using PGSs.

$$f\left(G_{ij}\right) \approx \sum_{z=1}^{n} \alpha^z g_{ij}^z = PGS_{ij}^{\mathrm{D}}$$

$$f\left(G_j\right) \approx \sum_{z=1}^{n} \delta^z \, g_j^z = PGS_j^{\mathrm{N}}$$

(3b.i)

α^z: True causal effect of a one allele change at i's genetic loci z on Y_{ij}

δ^z: True causal effect of a one allele change at either parent in family j's genetic loci z on Y_{ij}

g_{ij}^z: Total number of risk alleles at i's genetic loci z $(0, 1, \text{or } 2)$

g_j^z: Total number of risk alleles at the parents in family j's genetic loci z $(0, 1, 2, 3 \text{ or } 4)$

PGS_{ij}^{D}: PGS constructed from the true causal linear effect of i's genes on Y_{ij}

PGS_j^{N}: PGS constructed from the true causal linear effect of the parents in family j's genes on i's child$'$s Y_{ij}

Notice that $\vec{\alpha}$ is the vector of causal allelic weights used to construct the true, underlying PGS for direct genetic effects. In the same vein, $\vec{\delta}$ is the vector of causal allelic weights used to construct the true, underlying PGS for genetic nurture effects. Note that both $\vec{\alpha}$ and $\vec{\delta}$ are structural parameters that are never empirically observed.

In Equation 3b.ii, we rewrite our theoretical model using PGSs.

bioRxiv preprint doi: https://doi.org/10.1101/524850; this version posted July 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

$$Y_{ij} = \beta_0 + \beta_1 PGS_{ij}^{\mathrm{D}} + \beta_2 PGS_j^{\mathrm{N}} + f(E_{ij}) + f(E_j)$$

(3b.ii)

Notice that, because PGS_{ij}^{D} and PGS_j^{N} are not standardized, an individual's value for PGS_{ij}^{D} and PGS_j^{N} represents the true effect of their genes and their parents' genes, respectively, on Y_{ij} *in the units of* Y_{ij}. Thus, β_1 and β_2 are both equal to 1 by construction and Equation 3b.ii can equivalently be written without the β_1 and β_2 terms.

$$Y_{ij} = \beta_0 + PGS_{ij}^{\mathrm{D}} + PGS_j^{\mathrm{N}} + f(E_{ij}) + f(E_j)$$

(3b.iii)

3c. Transmitted Genetic Nurture Alleles

The presence of social genetic effects, such as genetic nurture effects, will only bias GWAS estimates of direct genetic effects when a social or biological process induces a correlation between the genetics of an individual and the genetics of his or her relevant social relationships. In the case of genetic nurture effects, biological recombination acts as such a process; children randomly inherit a portion of each parent's genome, leading to a mechanical correlation between parental genetics and child genetics. To capture the portion of the genetic nurture PGS that was transmitted to individual i in family j from their parents, we introduce a third PGS parameter that is absent from our formal model, PGS_{ij}^{N}.

$$\sum_{z=1}^{n} \delta^z\, g_{ij}^z = PGS_{ij}^{\mathrm{N}}$$

(3c.i)

PGS_{ij}^{N}: PGS constructed from the true causal linear effect of i's genes on i's child's Y

PGS_{ij}^{N} is constructed using aspects of both PGS_{ij}^{D} and PGS_j^{N}, the two PGSs corresponding to the two causal sources of genetic effects present in our theoretical model. Like PGS_{ij}^{D}, PGS_{ij}^{N} is constructed using g_{ij} (as opposed to g_j) and therefore varies within-families. However, like PGS_j^{N}, PGS_{ij}^{N} is constructed using the allelic weights $\vec{\delta}$, which correspond to genetic nurture effects (as opposed to direct genetic effects).

Notice that the relationship PGS_{ij}^{N} and PGS_j^{N} hinges on the relationship between g_{ij} and g_j. Because the alleles transmitted from parent (g_j) to child (g_{ij}) are determined stochastically through genetic recombination, the expected value of the correlation between g_{ij} and g_j is known. In Section A4 of the appendix, we derive the expected value of the association between PGS_{ij}^{N} and GS_j^{N} .

bioRxiv preprint doi: https://doi.org/10.1101/524850; this version posted July 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

$$\mathrm{E}\left[\rho_{PGS_{ij}^{\mathrm{N}},PGS_{j}^{\mathrm{N}}}\right] = \frac{\sqrt{2}}{2}$$

(3c.ii)

Note that this quantity does not vary between traits; this is due to the fact that same vector of allelic weights is used to construct both PGS_{ij}^{N} and PGS_{j}^{N} and differences are due entirely to the trait-independent recombination of parental alleles.

3d. The Relationship Between Direct Genetic Effects and Genetic Nurture Effects

While PGS_{ij}^{N} is absent from our underlying theoretical model, it provides the link between PGS_{ij}^{D} and PGS_{j}^{N} that allows for the presence of genetic nurture effects to distort GWAS results and subsequent PGS analysis. Thus, the most important components of our theoretical model are relationships between PGS_{ij}^{D} and PGS_{ij}^{N} and between PGS_{ij}^{N} and PGS_{j}^{N}. As we have seen above, PGS_{ij}^{N} and PGS_{j}^{N} have a mechanical correlation that does not vary between traits as it depends only on genotype (and not allelic weights). However, any differences between PGS_{ij}^{D} and PGS_{ij}^{N} are a result of differences between allelic weights used to construct each PGS ($\vec{\alpha}$ and $\vec{\delta}$, respectively). Because $\vec{\alpha}$ and $\vec{\delta}$ vary across traits, the extent to which PGS_{ij}^{D} and PGS_{ij}^{N} are correlated also varies between traits. We term the correlation between PGS_{ij}^{D} and PGS_{ij}^{N} for a given trait the *direct-nurture genetic correlation*.

$$\rho_g = \frac{cov\left(PGS_{ij}^{\mathrm{D}}, PGS_{ij}^{\mathrm{N}}\right)}{var\left(PGS_{ij}^{\mathrm{D}}\right)^{\frac{1}{2}} var\left(PGS_{ij}^{\mathrm{N}}\right)^{\frac{1}{2}}}$$

(3d.i)

ρ_g: Correlation between genetic nurture effects and direct genetic effects

In general, we expect the direct-nurture genetic correlation to be positive. Consider, for example, the case of educational attainment. Presumably some genetic pathways that lead a parent to create an environment conducive to their child succeeding in school will also have impacted the parent's schooling. However, there may also exist some genetic pathways that contribute to a parent's ability to create a positive educational environment for their children that do not influence the amount of educational attainment that the parent receives themselves. In the aggregate, we thus expect direct genetic effects and genetic nurture effects to have a correlation of less than 1. This value could in fact be negative, but, given the existing evidence suggests that ρ_g is typically positive (Kong et al. 2018; Bates et al. 2018; Belsky et al. 2018; Wertz et al. 2018), we focus on the case where the direct-nurture genetic correlation is bounded by 0 and 1.

3e. Underlying Allelic Weights

bioRxiv preprint doi: https://doi.org/10.1101/524850; this version posted July 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

We now turn to a discussion of α^z and δ^z. In the course of this discussion, we will show that ρ_g, a trait's direct-nurture genetic correlation, is a critical structural parameter of our theoretical model. We first discuss α^z. Allelic weights α^n $(n \in \{1 \dots N\})$ are taken from a distribution with variance σ. Without loss of generality, we define "risk" allele at each genetic loci z such that this distribution has mean 0. Turning to δ^z, we assume that $\vec{\delta}$ are drawn from an identical distribution as $\vec{\alpha}$, except with variance $\frac{\lambda^2}{4}\sigma$. Note that this is related to the variance of the $\vec{\alpha}$ distribution but scaled by a parameter λ; this allows for the average effects sizes to differ between direct genetic effects and genetic nurture effects. Finally, we assume that genetic nurture effects and genetic nature effects have a similar genetic architecture such that $\vec{\alpha}$ are $\vec{\delta}$ distributed in a similar way across genetic loci with respect to the allele frequency mean and variance. See Section A6 of the appendix for a formal treatment of this assumption.

The assumptions of our theoretical model entail that λ represents the ratio of the SNP heritabilities of genetic nurture effect and direct genetic effects (see Section A7 of the appendix). We call this value the *direct-nurture heritability ratio*.

$$\lambda = \mathrm{E}\left[\frac{2\sum_{z=1}^{n}(\delta^z \overline{g_{ij}^z})}{\sum_{z=1}^{n}(\alpha^z \overline{g_{ij}^z})}\right] = \frac{h_N^2}{h_D^2}$$

(3e.i)

λ: Ratio of the SNP heritabilities of genetic nurture effects and direct genetic effects

h_D^2: SNP heritability for direct genetic effects of Y_{ij}

h_N^2: SNP heritability for genetic nurture effects of Y_{ij}

Without loss of generality, we normalize all variables such that the variance of PGS_{ij}^{D} is equal to one.

$$var\left(PGS_{ij}^{\mathrm{D}}\right) = 1$$

(3e.ii)

We can now use our theoretical model to derive the expectation of the variance of the remaining PGSs as a function of the direct-nurture heritability ratio (see Sections A8 and A9 of the appendix).

$$\mathrm{E}\left[var\left(PGS_{ij}^{\mathrm{N}}\right)\right] = \frac{\lambda^2}{4}$$

bioRxiv preprint doi: https://doi.org/10.1101/524850; this version posted July 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

$$\mathrm{E}\left[var\left(PGS_j^{\mathrm{N}}\right)\right] = \frac{\lambda^2}{2}$$

(3e.iii)

Now that we have fully identified our theoretical model, we can use it to gain insights into the implications of genetic nurture effects for GWAS and PGSs.

4. Analytic Results

4a. Observed PGS

Up until this point, all our work has been theoretical; we have defined the functional form of a set of causal relationships between underlying parameters of interest which are difficult to observe directly. We now transport our theoretical model into the messy real world and consider its implications for GWAS and subsequent PGSs. In reality, we observe not $\vec{\alpha}$ but $\hat{\vec{\alpha}}$. We then use this observed $\hat{\vec{\alpha}}$ to construct not PGS_{ij}^{D} but $\widehat{PGS}_{ij}^{\mathrm{D}}$, which, as we will see, contains information from both PGS_{ij}^{D} and PGS_{ij}^{N}.

We obtain $\hat{\vec{\alpha}}$ by fitting the following regression for n SNPs via GWAS.

$$Y_{ij} = \hat{\beta}_0^z + \hat{\alpha}^z g_{ij}^z + \overrightarrow{X_{ij}}\hat{\vec{\Theta}} + \epsilon_{ij}$$

(4a.i)

$\hat{\alpha}^z$: Allelic weight from the observed linear relationship of a one allele change at $i's$ z^{th} gene and Y_{ij}

We can now plug in from our theoretical model (3b.iii).

$$\mathrm{E}[\hat{\alpha}^z] = \frac{var\left(g_{ij}^z\right)\alpha^z + cov\left(g_{ij}^z, g_j^z\right)\delta^z}{var\left(g_{ij}^z\right)}$$

(4a.ii)

To further analyze this expectation, we can separate g_j^z into the sum of alleles that were transmitted to i and the alleles that were not transmitted.

$$g_j^z{=}g_{ij}^z + f_{ij}^z$$

(4a.iii)

f_{ij}^z: Total number of risk alleles at the parents in family $j^{'}$s genetic loci z that were **not** transmitted to i $(0, 1, \text{or } 2)$

Thus yielding:

$$\mathrm{E}[\hat{\alpha}^z] = \frac{var\left(g_{ij}^z\right)\alpha^z + cov\left(g_{ij}^z, g_{ij}^z + f_{ij}^z\right)\delta^z}{var\left(g_{ij}^z\right)}$$

bioRxiv preprint doi: https://doi.org/10.1101/524850; this version posted July 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

$$\mathrm{E}[\hat{\alpha}^z] = \frac{var\left(g_{ij}^z\right)\alpha^z + cov\left(g_{ij}^z, g_{ij}^z\right)\delta^z + cov\left(g_{ij}^z, f_{ij}^z\right)\delta^z}{var\left(g_{ij}^z\right)}$$

(4a.iv)

In the absence of assortative mating, transmitted alleles are uncorrelated with non-transmitted allele, meaning that $cov\left(g_{ij}^z, f_{ij}^z\right) = 0$.

$$\mathrm{E}[\hat{\alpha}^z] = \frac{var\left(g_{ij}^z\right)\alpha^z + cov\left(g_{ij}^z, g_{ij}^z\right)\delta^z}{var\left(g_{ij}^z\right)}$$

$$\mathrm{E}[\hat{\alpha}^z] = \frac{var\left(g_{ij}^z\right)\alpha^z + var\left(g_{ij}^z\right)\delta^z}{var\left(g_{ij}^z\right)}$$

$$\mathrm{E}[\hat{\alpha}^z] = \alpha^z + \delta^z$$

(4a.v)

We can then use our observed vector of allelic weights for genetic nurture, $\vec{\hat{\alpha}}$, to construct our observed genetic nurture PGS, $\widehat{PGS}_{ij}^{\mathrm{D}}$.

$$\widehat{PGS}_{ij}^{\mathrm{D}} = \sum_{z=1}^{n} \hat{\alpha}^z \, g_{ij}^z$$

$$\widehat{PGS}_{ij}^{\mathrm{D}} = \sum_{z=1}^{n} (\alpha^z + \delta^z) \, g_{ij}^z$$

$$\widehat{PGS}_{ij}^{\mathrm{D}} = \ PGS_{ij}^{\mathrm{D}} + PGS_{ij}^{\mathrm{N}}$$

(4a.vi)

Finally, constructed PGSs are typically normalized within sample, so we convert $\widehat{PGS}_{ij}^{\mathrm{D}}$ to $\widehat{PGS'}_{ij}^{\mathrm{D}}$.

$$\widehat{PGS'}_{ij}^{\mathrm{D}} = \frac{\widehat{PGS}_{ij}^{\mathrm{D}} - \overline{\widehat{PGS}_{ij}^{\mathrm{D}}}}{var\left(\widehat{PGS}_{ij}^{\mathrm{D}}\right)^{\frac{1}{2}}}$$

bioRxiv preprint doi: https://doi.org/10.1101/524850; this version posted July 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

$$\widehat{PGS'}^{\mathrm{D}}_{ij} = \frac{\left(PGS^{\mathrm{D}}_{ij} + PGS^{\mathrm{N}}_{ij}\right) - \left(\overline{PGS}^{\mathrm{D}}_{ij} + \overline{PGS}^{\mathrm{N}}_{ij}\right)}{var\left(PGS^{\mathrm{D}}_{ij} + PGS^{\mathrm{N}}_{ij}\right)^{\frac{1}{2}}}$$

(4a.vii)

Thus, we have shown that the PGS derived from GWAS has information from both an individual's PGS for direct genetic effects and genetic nurture effects. In particular, note that in (4a.v) the estimated weights intermingle both α and δ. As a result, the quantity typically used for analysis, $\widehat{PGS}^{\mathrm{D}}_{ij}$, contains information about both the direct genetic effect and the genetic nurture effect (4a.vi).

4b. Between-Family Analyses

Our theoretical model has previously demonstrated that PGSs constructed from GWAS capture both the direct genetic effects and the genetic nurture effects of a given allele. We now explore the implications of this result for analyses using such PGSs, beginning with the between-family analysis (Model 1). We will see that the inclusion of genetic nurture effects biases direct genetic effect coefficients (away from zero).

$$\text{Model 1: } Y_i = \hat{\psi}_0 + \hat{\psi}_1 \widehat{PGS'}^{\mathrm{D}}_{ij} + \overrightarrow{X_{ij}}\overrightarrow{\hat{\Theta}} + \hat{\epsilon}_{ij}$$

(4b.i)

Now that we have derived $\widehat{PGS'}^{\mathrm{D}}_{ij}$, we can calculate the expected bias in $\hat{\psi}_1$. Recall that, because PGS^{D}_{ij} is unstandardized in our theoretical model (3b.iii), $\psi_1$and π_1 are mechanically both equal to one. Thus, the expected value of $\mathrm{E}\left[\hat{\psi}_1\right]$ and $\mathrm{E}[\hat{\pi}_1]$ represents the expected inflation or deflation of estimated PGS coefficients (i.e. bias) in between-family and within-family analyses, respectively.

$$\mathrm{E}\left[\hat{\psi}_1\right] = \frac{cov\left(\widehat{PGS'}^{\mathrm{D}}_{ij}, Y_{ij}\right)}{var\left(\widehat{PGS'}^{\mathrm{D}}_{ij}\right)}$$

(4b.ii)

Note that $cov\left(\widehat{PGS'}^{\mathrm{D}}_{ij}, Y_{ij}\right)$ is given by true causal relationships from our theoretical model (3b.iii).

$$\mathrm{E}\left[\hat{\psi}_1\right] = \frac{cov\left(\widehat{PGS'}^{\mathrm{D}}_{ij}, \widehat{PGS'}^{\mathrm{D}}_{ij}\right)\beta_1 + cov\left(\widehat{PGS'}^{\mathrm{D}}_{ij}, \widehat{PGS'}^{\mathrm{D}}_{ij}\right)\beta_2}{var\left(\widehat{PGS'}^{\mathrm{D}}_{ij}\right)}$$

(4b.iii)

bioRxiv preprint doi: https://doi.org/10.1101/524850; this version posted July 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

β_1 and β_2 from our theoretical model are equal to 1 by construction and therefore fall away.

$$\mathrm{E}[\hat{\psi}_1] = \frac{cov\left(\widehat{PGS'}^{\mathrm{D}}_{ij}, \widehat{PGS'}^{\mathrm{D}}_{ij}\right) + cov\left(\widehat{PGS'}^{\mathrm{D}}_{ij}, \widehat{PGS'}^{\mathrm{D}}_{ij}\right)}{var\left(\widehat{PGS'}^{\mathrm{D}}_{ij}\right)}$$

(4b.iv)

We solve (see Section A9 of the appendix for details) to obtain the magnitude of the bias in $\hat{\psi}_1$.

$$\mathrm{E}[\hat{\psi}_1] = \sqrt{1 + \lambda\rho_g + \frac{\lambda^2}{4}}$$

(4b.v)

So we have shown that our observed estimates for $\hat{\psi}_1$ will be biased upwards by a factor of $\sqrt{1 + \lambda\rho_g + \frac{\lambda^2}{4}}$. We will unpack this quantity further in the discussion but note that it depends on unobserved parameters.

4c. Within-Family Analyses

Let's now turn to our within-family analysis (Model 2). We will see that the inclusion of genetic nurture effects biases direct genetic effect coefficients towards zero.

$$\text{Model 2: } Y_{ij} = \hat{\pi}_0 + \hat{\pi}_1 \widehat{PGS'}^{\mathrm{D}}_{ij} + \overrightarrow{X_{ij}}\overrightarrow{\hat{\Theta}} + \Gamma_{\mathrm{j}} + \hat{\epsilon}_{ij}$$

Or, equivalently:

$$\left(Y_{1j} - Y_{0j}\right) = \hat{\pi}_1\left(\widehat{PGS'}^{\mathrm{D}}_{1j} - \widehat{PGS'}^{\mathrm{D}}_{0j}\right) + \left(\hat{\epsilon}_{1j} - \hat{\epsilon}_{0j}\right)$$
$$\Delta_0^1 Y_{ij} = \hat{\pi}_1 \Delta_0^1 \widehat{PGS'}^{\mathrm{D}}_{ij} + \Delta_0^1 \hat{\epsilon}_{ij}$$

(4c.i)

PGS^{N}_{0j}: PGS^{N}_{ij} of sibling 0 in family j

PGS^{N}_{1j}: PGS^{N}_{ij} of sibling 1 in family j

As before, we begin by deriving the expected value of $\hat{\pi}_1$.

$$\mathrm{E}[\hat{\pi}_1] = \frac{cov\left(\Delta_0^1 \widehat{PGS'}^{\mathrm{D}}_{ij}, \Delta_0^1 Y_{ij}\right)}{var\left(\Delta_0^1 \widehat{PGS'}^{\mathrm{D}}_{ij}\right)}$$

(4c.ii)

bioRxiv preprint doi: https://doi.org/10.1101/524850; this version posted July 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

$cov\left(\Delta_0^1 \widehat{PGS'}_{ij}^{\mathrm{D}}, \Delta_0^1 Y_{ij}\right)$ is given by true causal relationships from our theoretical model.

$$\mathrm{E}[\hat{\pi}_1] = \frac{cov\left(\Delta_0^1 \widehat{PGS'}_{ij}^{\mathrm{D}}, \Delta_0^1 PGS_{ij}^{\mathrm{D}}\right)\beta_1 + cov\left(\Delta_0^1 \widehat{PGS'}_{ij}^{\mathrm{D}}, \Delta_0^1 PGS_j^{\mathrm{N}}\right)\beta_2}{var\left(\Delta_0^1 \widehat{PGS'}_{ij}^{\mathrm{D}}\right)}$$

(4c.iii)

Again, β_1 and β_2 from our theoretical model are equal to 1 by construction and therefor fall away.

$$\mathrm{E}[\hat{\pi}_1] = \frac{cov\left(\Delta_0^1 \widehat{PGS'}_{ij}^{\mathrm{D}}, \Delta_0^1 PGS_{ij}^{\mathrm{D}}\right) + cov\left(\Delta_0^1 \widehat{PGS'}_{ij}^{\mathrm{D}}, \Delta_0^1 PGS_j^{\mathrm{N}}\right)}{var\left(\Delta_0^1 \widehat{PGS'}_{ij}^{\mathrm{D}}\right)}$$

(4c.iv)

Notice that between siblings there is no variation in family genetic nurturing environment, meaning that $\Delta_0^1 PGS_j^{\mathrm{N}} = 0$.

$$\mathrm{E}[\hat{\pi}_1] = \frac{cov\left(\Delta_0^1 \widehat{PGS'}_{ij}^{\mathrm{D}}, \Delta_0^1 PGS_{ij}^{\mathrm{D}}\right)}{var\left(\Delta_0^1 \widehat{PGS'}_{ij}^{\mathrm{D}}\right)}$$

(4c.v)

Now we just solve (see Section A11 of the appendix for details) to obtain the magnitude of the bias in $\hat{\pi}_1$.

$$\mathrm{E}[\hat{\pi}_1] = \frac{1 + \frac{\lambda\rho_g}{2}}{\sqrt{1 + \lambda\rho_g + \frac{\lambda^2}{4}}}$$

(4c.vi)

Recall that, because PGS_{ij}^{D} is unstandardized in our theoretical model (3b.iii), π_1 is mechanically equal to one. Thus, the expected value of $\hat{\pi}_1$ represents inflation or deflation of PGS coefficient estimates and is itself interpretable as measure of bias. Thus, we have shown that our observed estimates for $\hat{\psi}_1$ will be biased downwards by a factor of $\frac{1+\frac{\lambda\rho_g}{2}}{\sqrt{1+\lambda\rho_g+\frac{\lambda^2}{4}}}$. We discuss this quantity further in the discussion.

5. Discussion

5a. Bias

bioRxiv preprint doi: https://doi.org/10.1101/524850; this version posted July 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Our theoretical model illustrates that, in the presence of genetic nurture (i.e. $h_N^2 \neq 0$), regression analyses using PGSs to estimate the effects of an individual's genetics on their outcomes will suffer from bias. Between-family OLS models will be biased upwards by a factor of $\sqrt{1 + \lambda\rho_g + \frac{\lambda^2}{4}}$ while within-family fixed effect models will be biased downwards by a factor of $\frac{1+\frac{\lambda\rho_g}{2}}{\sqrt{1+\lambda\rho_g+\frac{\lambda^2}{4}}}$. Figure 1 plots the bias in both OLS and family fixed effect regressions using PGSs as a function of various direct-nurture genetic correlations and direct-nurture heritability ratios.

[Insert Figure 1 Here]

The absence of bias is represented in Figure 1 by the horizontal line at y=1. For any given trait, bias is always larger in the between-family models. The magnitude of bias is a function of two parameters: ρ_g, the trait's direct-nurture genetic correlation, and λ, the trait's direct-nurture heritability ratio. As λ increases, so does the magnitude of the bias. However, ρ_g has opposing effects on the bias within and between-families; as ρ_g increases, we note more downward within-family bias but more upward between-family bias. Thus, taken together, $\hat{\psi}_1$ and $\hat{\pi}_1$ can provide a useful set of upper and lower bounds of the true causal effects.

There is an intuitive interpretation to the trends presented in Figure 1. In the between-family estimates, the inclusion of genetic nurture effects in $\widehat{PGS}_{ij}^{\mathrm{D}}$ leads to an upward bias as it is capturing both differences in genetic composition between individuals *and* differences in the family environments between individuals that result from differences in their parents' genetic composition. The larger that ρ_g is, the greater extent to which an individual with a beneficial allele for educational attainment reaps the reward from the same genes twice; first when their parents provide a more nurturing environment, and second when they themselves inherit the beneficial allele.

On the other hand, in family fixed effects models (within-family), the inclusion of genetic nurture in the observed PGS leads to downward bias when ρ_g is less than unity. This is because there are no differences in genetic nurture systematically driving differences in educational outcomes between siblings. Regardless of their own PGS_{ij}^{D}, siblings have identical PGS_j^{N} because they are in the same family j. Thus, any extent that genetic nurture effects causes $\hat{\vec{\alpha}}$ to diverge from $\vec{\alpha}$ amounts to measurement error in the allelic weights and causes downward attenuation bias.

The insights from our theoretical model offer a partial explanation of the large differences in $\frac{\hat{\pi}_1}{\hat{\psi}_1}$ observed across the various phenotypes considered in Table 1. For traits like years of schooling, cognitive ability, and (more speculatively) depression, where indirect genetic effects likely play an important role (i.e. large λ) , we see large differences between $\hat{\psi}_1$ and $\hat{\pi}_1$

bioRxiv preprint doi: https://doi.org/10.1101/524850; this version posted July 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(i.e. $\frac{\hat{\pi}_1}{\hat{\psi}_1} < 1$). On the other side of the coin, for traits like body mass index and height, where most of the genetic contribution is likely to be direct (i.e. small λ) , we see almost no difference between $\hat{\psi}_1$ and $\hat{\pi}_1$ (i.e. $\frac{\hat{\pi}_1}{\hat{\psi}_1} \approx 1$). Turning back to Figure 1, you can see this coincides with what our theoretical model would have predicted. Further, our model suggests that the differences in $\frac{\hat{\pi}_1}{\hat{\psi}_1}$ observed between years of schooling, cognitive ability, and depression may be a function of differences in ρ_g between the traits.

Recall that these theoretical results are based on several simplifying assumptions: no gene-environment interaction, no gene-environment correlation, and no assortative mating. These three assumptions are unlikely to hold for most complex traits studied of interest to the social and biological sciences. Nonetheless, we can use the results from this simplified model to begin to probe how such violations might influence our results. Consider the case of positive genetic assortative mating, which exists for many of the traits considered in this paper (Yengo, Robinson, et al. 2018). A correlation between maternal and paternal genetics induces a positive correlation between transmitted and non-transmitted parental alleles shown in (4a.iv). This is because non-transmitted maternal alleles would be correlated with transmitted paternal alleles and vice versa. Thus, $cov\left(g_{ij}^{z}, f_{ij}^{z}\right)$ is no longer equal to zero and does not fall out of our equation. In such a case, the expected value of a GWAS allelic weight, shown in (4a.v), would include information about both non-transmitted genetic nurture effects in addition to the direct genetic effects and the transmitted genetic nurture effects. Thus, the bias documented in PGS analyses will increase as positive genetic assortative mating increases.

Next, we consider the case of gene-environment correlation. It is important to note that the bias that genetic nurture may cause in GWAS and PGS results from a special case of gene-environment correlation (i.e. a gene-environment correlation that mechanically exists due to the correlation of genetics between parents and their children induced though genetic inheritance). Thus, in our theoretical model's specification, gene-environment correlation only exists when environmental features relevant to an outcome are correlated with an individual's genetics after accounting for their parents' genetics. When such a gene-environment correlation exists, GWAS results and PGSs become biased in exactly the same way as they do from genetic nurture effects alone. For example, the existence of positive gene-environment correlation effectively increases λ (as additional outcome variance explained by a non-genetic nurture environmental component), therein increasing the magnitude of the bias.

Finally, the case of gene-environment interaction is difficult to consider more generally, as it would vary as a function of the magnitude, direction, and the pathways of the interaction. Thus, the effects of gene-environment interactions on how genetic nurture effects influence GWAS and PGSs remain uncertain.

5b. Direct-Nurture Genetic Correlation and Heritability Ratio

bioRxiv preprint doi: https://doi.org/10.1101/524850; this version posted July 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Numerous GWAS have been conducted in the last decade (Visscher et al. 2017; Mills and Rahal 2019). Nonetheless, to our knowledge, no GWAS has been conducted in human populations that independently identifies the direct genetic effects and the genetic nurture effects for a complex trait (a notable exception is GWAS on maternal influences on child birthweight (Beaumont et al. 2018; Warrington et al. 2019), though this indirect genetic affect does not seem to be social in nature). Thus, for virtually all complex traits, little is known about the parameters of interest identified in our models: the direct-nurture genetic correlation and the direct-nurture heritability ratio (ρ_g and λ respectively). Critically, the two parameters are readily estimable with existing data and methods (which we discuss further in 5d).

A better understanding of ρ_g and λ would offer value beyond aiding in the comparison of results from within-family and between-family regressions. The genetic pathways discovered in GWASs and summarized in PGSs offer researchers a puzzle to unpack (Freese 2018). Understanding why some phenotypes have strong versus weak social genetic effects, or why the link between direct genetic and social genetic influence are more versus less linked, could help researchers glean insight into the underlying mechanisms at play.

Moreover, it would be interesting to understand how ρ_g and λ are influenced by the social environment. For example, while gene-environment interaction studies have been the conventional way to understand how the environment moderates the influence of genetics, recent work has proposed a genetic correlation-environment interaction study (Wedow et al. 2018). In a genetic correlation-environment interaction study framework, the social environment can transform the genetic link between two traits. Exploring how the environment shapes $\rho_g$would be a special case of a genetic correlation-environment interaction study where the two traits influence the same phenotype (directly and socially). Social policymakers might prefer a low ρ_g for valued life outcomes like educational attainment to reduce the accumulation of inequality across generations.

While the specific social, physical, or economic factors that moderate ρ_g and λ for various traits remains to be explored empirically, there may exist *a priori* reasons to suspect certain environmental modifiers. For example, let us say that individual variation in height is a function of both direct genetic effects that shape physiological development and genetic nurture effects that influence access to socioeconomic and nutritional resources during childhood (thereby reducing the likelihood of stunting). If there is a large casual effect of height on socioeconomic status, we'd expect individuals with a greater genetic predisposition for height (i.e. a high $\widehat{PGS}_{ij}^{\mathrm{D}}$) to be more likely to attain a higher socioeconomic position where their children to have access to nutritional and health resources (i.e. a high $\widehat{PGS}_{ij}^{\mathrm{N}}$) resulting in a positive ρ_g. However, this relationship could be modified by environmental features; if, for instance, the causal effect of height on social status is due to labor force discrimination, outlawing the use of height for employment decisions would uncouple $\widehat{PGS}_{ij}^{\mathrm{D}}$ and $\widehat{PGS}_{ij}^{\mathrm{N}}$ and reduce ρ_g. Alternatively, a social policy that provides adequate healthcare and nutrition to all children could effectively eliminate

bioRxiv preprint doi: https://doi.org/10.1101/524850; this version posted July 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

stunting and undo the relationship between parental genetics for height altogether, forcing both ρ_g and λ to zero.

5c. Implications for the Use of Polygenic Scores

While within-family analyses have demonstrated that many PGSs do have significant casual signal for direct genetic effects, comparing the results from within-family analyses to results from between-family analyses is complicated by the presence of genetic nurture effects. To what extent do existing PGSs capture direct genetic effects, genetic nurture effects, and socioeconomic or geographic confounding? Until we better understand ρ_g and λ for a wide variety of traits, our ability to use within-family analyses to validate between-family discoveries will be limited. Analyses using PGSs should be interpreted accordingly.

Even in the absence of confounding due to population stratification, the observation that existing PGSs likely have genetic nurture components complicates their use and interpretation. Say, for instance, a researcher wonders whether there exists moderation of the association between an individual's PGS and their educational attainment as a function of school-level socioeconomic status (Trejo et al. 2018). Because a component of the PGS is capturing the benefit of having a parent with higher educational attainment and (in turn a higher socioeconomic status), any detected gene-environment interaction lacks a clear interpretation. It might be the case that household socioeconomic status interacts with school-level socioeconomic status in shaping educational attainment, or alternatively it could be that an individual's genetic composition interacts with school-level socioeconomic status in shaping educational attainment. These two different results have very different theoretical and practical implications but are indistinguishable in analyses using existing PGSs, which contain both direct genetic effects and genetic nurture effects.

5d. Future Research

The models constructed in this paper highlight key areas for future research in the field of social science genomics. Across a range of complex phenotypes, there is much work to be done towards separating out the genetics nurture effects from direct genetic effects. Utilizing random cage-mate assignment in mice, a recent study in mice conducted a social genetic effects GWAS and direct genetic effects GWAS in parallel and identified statistically significant genome wide social genetic effect loci for 16 phenotypes (Baud et al. 2018). For these 16 phenotypes, the mean social-direct genetic correlation was 0.53 and the mean social-direct heritability ratio was 1.29. Crucially, social genetic effects arise from partially different loci as direct genetic effects and can have effects of differing magnitudes or directions at the same loci.

Unfortunately, social relationships are often not randomly assigned in human populations. Nonetheless, existing GWAS methods could be modified to use dyads of parents and their children. For example, a social genetic effects GWAS could be conducted by controlling for child's genetics in a GWAS of parental genetics on child phenotype. Alternatively, a GWAS conducted using variation only amongst sibling pairs would provide

bioRxiv preprint doi: https://doi.org/10.1101/524850; this version posted July 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

information on direct genetic pathways untainted by confounding or genetic nurture effects. Results from such a sibling GWAS might then be used to back out information about the genetic nurture effects from existing GWAS results of unrelated individuals. Nonetheless, obtaining large samples of parent-child or sibling pairs might prove challenging for many complex phenotypes. If statistical power is a problem, methods such as LD score regression (Bulik-Sullivan et al. 2015) could be used with smaller samples to identify estimates of ρ_g and λ. Indeed, having estimates of ρ_g and λ for a trait would allow researchers to correct for bias in between-family and within-family analyses that use PGSs by dividing observed regression coefficients by the quantities displayed in (4b.v) and (4c.vi), respectively.

It is also possible that the direct effects and social genetic effects are not independent. Imagine, for example, that parents with a higher educational attainment PGSs tend to invest more heavily in their children with higher PGSs than do parents with lower PGSs (differential investment in low birth weight children has been observed by across socioeconomic lines (Hsin 2012; Restrepo 2016)). In such a case, the effects of parental genetics on a child would vary as a function as their child's genetics. Research designs that investigate the existence of interaction between direct genetic effects and genetic nurture effects may prove a fruitful avenue for future inquiry.

Finally, work should be done to extend our framework to other social genetic effects within-families, such as between sibling pairs. We chose to start with genetic nurture effects because, unlike sibling effects, genetic nurture effects are likely to be unidirectional, with the causal effect pointing from parent to child, making them more straightforward to model (effects between siblings are likely reciprocal) (Kong et al. 2018) and because parental effects generalize to all families, not just those with multiple children. Nonetheless, social genetic effects between siblings may also complicate the interpretation of GWAS results and warrant attention.

5d. Conclusion

In summary, we formalize a theoretical model for additive direct genetic effects and genetic nurture effects and show that, unlike bias from other confounders, the presence of genetic nurture can bias coefficients from between-family and within-family regressions using PGSs. While within-family analyses that compare siblings using family fixed effects are considered the gold-standard, they are not without their own complications. Even if we were able to run a GWAS on an infinitely large sample using the methods in practice today, the presence of confounding social genetic effects would mean that it would be impossible to obtain precise estimates of the causal effect of an individual's genes on their life outcomes. Until GWAS can be conducted controlling for parental genotype, models using PGSs may be biased by genetic nurture effects. Obtaining estimates of a trait's direct-nurture genetic correlation and direct-nurture heritability ratio may allow researchers to correct for such a bias.

bioRxiv preprint doi: https://doi.org/10.1101/524850; this version posted July 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

References

Barcellos, Silvia H., Leandro S. Carvalho, and Patrick Turley. 2018. "Education Can Reduce Health Differences Related to Genetic Risk of Obesity." *Proceedings of the National Academy of Sciences*. https://doi.org/10.1007/BF00871674.

Bates, Timothy C., Brion S. Maher, Sarah E. Medland, Kerrie McAloney, Margaret J. Wright, Narelle K. Hansell, Kenneth S. Kendler, Nicholas G. Martin, and Nathan A. Gillespie. 2018. "The Nature of Nurture: Using a Virtual-Parent Design to Test Parenting Effects on Children's Educational Attainment in Genotyped Families." *Twin Research and Human Genetics*. https://doi.org/10.1017/thg.2018.11.

Baud, Amelie, Francesco Paolo Casale, Jerome Nicod, and Oliver Stegle. 2018. "Genome-Wide Association Study of Social Genetic Effects on 170 Phenotypes in Laboratory Mice." *BioRxiv*. https://doi.org/10.1101/302349.

Beaumont, Robin N., Nicole M. Warrington, Alana Cavadino, Jessica Tyrrell, Michael Nodzenski, Momoko Horikoshi, Frank Geller, et al. 2018. "Genome-Wide Association Study of Offspring Birth Weight in 86 577 Women Identifies Five Novel Loci and Highlights Maternal Genetic Effects That Are Independent of Fetal Genetics." *Human Molecular Genetics*. https://doi.org/10.1093/hmg/ddx429.

Belsky, D. W., T. E. Moffitt, D. L. Corcoran, B. Domingue, H. Harrington, S. Hogan, R. Houts, et al. 2016. "The Genetics of Success: How Single-Nucleotide Polymorphisms Associated With Educational Attainment Relate to Life-Course Development." *Psychological Science* 27:957–72. https://doi.org/10.1177/0956797616643070.

Belsky, Daniel W., Benjamin W. Domingue, Robbee Wedow, Louise Arseneault, Jason D. Boardman, Avshalom Caspi, Dalton Conley, et al. 2018. "Genetic Analysis of Social-Class Mobility in Five Longitudinal Studies." *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.1801238115.

Belsky, Daniel W., and Salomon Israel. 2014. "Integrating Genetics and Social Science: Genetic Risk Scores." *Biodemography and Social Biology*. https://doi.org/10.1080/19485565.2014.946591.

Belsky, Daniel W., Terrie E. Moffitt, Timothy B. Baker, Andrea K. Biddle, James P. Evans, Hona Lee Harrington, Renate Houts, et al. 2013. "Polygenic Risk and the Developmental Progression to Heavy, Persistent Smoking and Nicotine Dependence: Evidence from a 4-Decade Longitudinal Study." *JAMA Psychiatry*. https://doi.org/10.1001/jamapsychiatry.2013.736.

Bergsma, R., E. Kanis, E. F. Knol, and P. Bijma. 2008. "The Contribution of Social Effects to Heritable Variation in Finishing Traits of Domestic Pigs (Sus Scrofa)." *Genetics*. https://doi.org/10.1534/genetics.107.084236.

Bulik-Sullivan, Brendan K, Po-Ru Loh, Hilary K Finucane, Stephan Ripke, Jian Yang, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Nick Patterson, Mark J Daly, Alkes L Price, and Benjamin M Neale. 2015. "LD Score Regression Distinguishes Confounding from Polygenicity in Genome-Wide Association Studies." *Nat Genet* advance on (3). Nature Publishing Group:291–95. https://doi.org/10.1038/ng.3211.

Canario, L., N. Lundeheim, and P. Bijma. 2017. "The Early-Life Environment of a Pig Shapes the Phenotypes of Its Social Partners in Adulthood." *Heredity*. https://doi.org/10.1038/hdy.2017.3.

Cawley, John, Euna Han, Jiyoon (June) Kim, and Edward C. Norton. 2017. "Testing for Peer

bioRxiv preprint doi: https://doi.org/10.1101/524850; this version posted July 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Effects Using Genetic Data." *NBER Working Paper No. 23719*. https://doi.org/10.3386/w23719.

Chabris, C. F., J. J. Lee, D. Cesarini, D. J. Benjamin, and D. I. Laibson. 2015. "The Fourth Law of Behavior Genetics." *Current Directions in Psychological Science* 24 (4):304–12. https://doi.org/10.1177/0963721415580430.

Conley, Dalton, and Jason Fletcher. 2017. *The Genome Factor What the Social Genomics Revolution Reveals about Ourselves, Our History, and the Future*.

Domingue, Benjamin W., and Daniel W. Belsky. 2017. "The Social Genome: Current Findings and Implications for the Study of Human Genetics." *PLoS Genetics* 13 (3). https://doi.org/10.1371/journal.pgen.1006615.

Domingue, Benjamin W., Daniel W. Belsky, Jason M. Fletcher, Dalton Conley, Jason D. Boardman, and Kathleen Mullan Harris. 2018. "The Social Genome of Friends and Schoolmates in the National Longitudinal Study of Adolescent to Adult Health." *Proceedings of the National Academy of Sciences*, 201711803. https://doi.org/10.1073/pnas.1711803115.

Domingue, Benjamin W, Daniel W Belsky, Dalton Conley, Kathleen Mullan Harris, and Jason D Boardman. 2015. "Polygenic Influence on Educational Attainment: New Evidence From the National Longitudinal Study of Adolescent to Adult Health." *AERA Open* 1 (3):1–13. https://doi.org/10.1177/2332858415599972.

Dudbridge, Frank. 2013. "Power and Predictive Accuracy of Polygenic Risk Scores." *PLoS Genetics* 9 (3). https://doi.org/10.1371/journal.pgen.1003348.

Duncan, Laramie, Hanyang Shen, Bizu Gelaye, Kerry Ressler, Marcus Feldman, Roseann Peterson, and Benjamin Domingue. 2018. "Analysis of Polygenic Score Usage and Performance in Diverse Human Populations." *BioRxiv*. https://doi.org/10.1101/398396.

Freese, Jeremy. 2018. "The Arrival of Social Science Genomics." *Contemporary Sociology*. https://doi.org/10.1177/0094306118792214a.

Giri, Ayush, Jacklyn N Hellwege, Jacob M Keaton, Jihwan Park, Chengxiang Qiu, Helen R Warren, Eric S Torstenson, et al. 2019. "Trans-Ethnic Association Study of Blood Pressure Determinants in over 750,000 Individuals." *Nature Genetics* 51 (1). Nature Publishing Group:51.

Hamer, D. H., and L. Sirota. 2000. "Beware the Chopsticks Gene." *Molecular Psychiatry*. https://doi.org/10.1038/sj.mp.4000662.

Harris, Kathleen Mullan. 2013. "The Add Health Study: Design and Accomplishments." *Chapel Hill: Carolina Population Center, University of North Carolina at Chapel Hill*. https://doi.org/10.17615/C6TW87.

Hsin, Amy. 2012. "Is Biology Destiny ? Birth Weight and Differential Parental Treatment," 1385–1405. https://doi.org/10.1007/s13524-012-0123-y.

Hyde, Craig L., Michael W. Nagle, Chao Tian, Xing Chen, Sara A. Paciga, Jens R. Wendland, Joyce Y. Tung, David A. Hinds, Roy H. Perlis, and Ashley R. Winslow. 2016. "Identification of 15 Genetic Loci Associated with Risk of Major Depression in Individuals of European Descent." *Nature Genetics*. https://doi.org/10.1038/ng.3623.

Kong, Augustine, Gudmar Thorleifsson, Michael L. Frigge, Bjarni J. Vilhjalmsson, Alexander I. Young, Thorgeir E. Thorgeirsson, Stefania Benonisdottir, et al. 2018. "The Nature of Nurture: Effects of Parental Genotypes." *Science* 359 (6374):424–28. https://doi.org/10.1126/science.aan6877.

Lee, James J., Robbee Wedow, Aysu Okbay, Edward Kong, Omeed Maghzian, Meghan Zacher,

bioRxiv preprint doi: https://doi.org/10.1101/524850; this version posted July 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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*. https://doi.org/10.1038/s41588-018-0147-3.

Locke, Adam E., Bratati Kahali, Sonja I. Berndt, Anne E. Justice, Tune H. Pers, Felix R. Day, Corey Powell, et al. 2015. “Genetic Studies of Body Mass Index Yield New Insights for Obesity Biology.” *Nature*. https://doi.org/10.1038/nature14177.

Martschenko, Daphne, Sam Trejo, and Benjamin W. Domingue. 2019. “Genetics and Education: Recent Developments in the Context of an Ugly History and an Uncertain Future.” *AERA Open* 5 (1):1–15. https://doi.org/10.1093/ije/dyx041.

Mills, Melinda C., and Charles Rahal. 2019. “A Scientometric Review of Genome-Wide Association Studies.” *Communications Biology* 2 (1). Springer US:9. https://doi.org/10.1038/s42003-018-0261-x.

Moore, Allen J., Edmund D. Brodie, and Jason B. Wolf. 1997. “INTERACTING PHENOTYPES AND THE EVOLUTIONARY PROCESS: I. DIRECT AND INDIRECT GENETIC EFFECTS OF SOCIAL INTERACTIONS.” *Evolution*. https://doi.org/10.2307/2411187.

Novembre, John, Toby Johnson, Katarzyna Bryc, Zoltán Kutalik, Adam R. Boyko, Adam Auton, Amit Indap, et al. 2008. “Genes Mirror Geography within Europe.” *Nature*. https://doi.org/10.1038/nature07331.

Okbay, Aysu, Bart M.L. Baselmans, Jan Emmanuel De Neve, Patrick Turley, Michel G. Nivard, Mark Alan Fontana, S. Fleur W. Meddens, et al. 2016. “Genetic Variants Associated with Subjective Well-Being, Depressive Symptoms, and Neuroticism Identified through Genome-Wide Analyses.” *Nature Genetics*. https://doi.org/10.1038/ng.3552.

Pearson, Thomas a, and Teri a Manolio. 2008. “How to Interpret a Genome-Wide Association Study.” *JAMA : The Journal of the American Medical Association* 299 (11):1335–44. https://doi.org/10.1001/jama.299.11.1335.

Petfield, D, S F Chenoweth, H D Rundle, and Mark W Blows. 2005. “Genetic Variance in Female Condition Predicts Indirect Genetic Variance in Male Sexual Display Traits.” *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.0409378102.

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.

Restrepo, Brandon J. 2016. “Parental Investment Responses to a Low Birth Weight Outcome : Who Compensates and Who Reinforces ?” *Journal of Population Economics*. Journal of Population Economics, 969–89. https://doi.org/10.1007/s00148-016-0590-3.

Rietveld, Cornelius A., Sarah E. Medland, Jaime Derringer, Jian Yang, Tõnu Esko, Nicolas W. Martin, Harm Jan Westra, et al. 2013. “GWAS of 126,559 Individuals Identifies Genetic Variants Associated with Educational Attainment.” *Science*. https://doi.org/10.1126/science.1235488.

Rietveld, Cornelius A, Dalton Conley, Nicholas Eriksson, Tõnu Esko, Sarah E Medland, Anna A E Vinkhuyzen, Jian Yang, et al. 2014. “Replicability and Robustness of Genome-Wide-Association Studies for Behavioral Traits.” *Psychological Science* 25 (11):1975–86. https://doi.org/10.1177/0956797614545132.

Selzam, Saskia, Stuart J Ritchie, Jean-baptiste Pingault, Chandra A Reynolds, and F Paul. 2019.

bioRxiv preprint doi: https://doi.org/10.1101/524850; this version posted July 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

"Comparing Within- and between-Family Polygenic Score Prediction Authors," 1–32.

Sotoudeh, Ramina, Dalton Conley, and Kathleen Mullan Harris. 2017. "The Influence of Peer Genotypes and Behavior on Smoking Outcomes: Evidence from Add Health." *NBER Working Paper Series*, 51. https://doi.org/10.3386/w24113.

Trejo, Sam, Daniel Belsky, Jason Boardman, Jeremy Freese, Kathleen Harris, Pam Herd, Kamil Sicinski, and Benjamin Domingue. 2018. "Schools as Moderators of Genetic Associations with Life Course Attainments: Evidence from the WLS and Add Heath." *Sociological Science*. https://doi.org/10.15195/v5.a22.

Turley, Patrick, Raymond K. Walters, Omeed Maghzian, Aysu Okbay, James J. Lee, Mark Alan Fontana, Tuan Anh Nguyen-Viet, et al. 2018. "Multi-Trait Analysis of Genome-Wide Association Summary Statistics Using MTAG." *Nature Genetics*. https://doi.org/10.1038/s41588-017-0009-4.

Visscher, Peter M., Naomi R. Wray, Qian Zhang, Pamela Sklar, Mark I. McCarthy, Matthew A. Brown, and Jian Yang. 2017. "10 Years of GWAS Discovery: Biology, Function, and Translation." *American Journal of Human Genetics*. https://doi.org/10.1016/j.ajhg.2017.06.005.

Warrington, NM, RN Beaumont, M Horikoshi, FR Day, and O Helgeland. 2019. "Maternal and Fetal Genetic Effects on Birth Weight and Their Relevance to Cardio-Metabolic Risk Factors." *Nature Genetics*, no. Ld. https://doi.org/10.1038/s41588-019-0403-1.

Wedow, Robbee, Meghan Zacher, Brooke M. Huibregtse, Kathleen Mullan Harris, Benjamin W. Domingue, and Jason D. Boardman. 2018. "Education, Smoking, and Cohort Change: Forwarding a Multidimensional Theory of the Environmental Moderation of Genetic Effects." *American Sociological Review*. https://doi.org/10.1177/0003122418785368.

Wertz, Jasmin, Terrie E Moffitt, Jessica Agnew-Blais, Louise Arseneault, Daniel W. Belsky, David. L. Corcoran, Renate Houts, et al. 2018. "Using DNA from Mothers and Children to Study Parental Investment in Children's Educational Attainment." *BioRxiv*.

Wolf, Jason B., Edmund D. Brodie, James M. Cheverud, Allen J. Moore, and Michael J. Wade. 1998. "Evolutionary Consequences of Indirect Genetic Effects." *Trends in Ecology and Evolution*. https://doi.org/10.1016/S0169-5347(97)01233-0.

Wood, Andrew R., Tonu Esko, Jian Yang, Sailaja Vedantam, Tune H. Pers, Stefan Gustafsson, Audrey Y. Chu, et al. 2014. "Defining the Role of Common Variation in the Genomic and Biological Architecture of Adult Human Height." *Nature Genetics*. https://doi.org/10.1038/ng.3097.

Yengo, Loic, Matthew R Robinson, Matthew C Keller, Kathryn E Kemper, Yuanhao Yang, Maciej Trzaskowski, Jacob Gratten, et al. 2018. "Imprint of Assortative Mating on the Human Genome," 1–41.

Yengo, Loic, Julia Sidorenko, Kathryn E. Kemper, Zhili Zheng, Andrew R. Wood, Michael N. Weedon, Timothy M. Frayling, Joel Hirschhorn, Jian Yang, and Peter M. Visscher. 2018. "Meta-Analysis of Genome-Wide Association Studies for Height and Body Mass Index in ∼700000 Individuals of European Ancestry." *Human Molecular Genetics*. https://doi.org/10.1093/hmg/ddy271.

Tables

Table 1. The association between polygenic score and observed trait for six phenotypes, within-families and between-families.

	$\hat{\psi}_1$	$\hat{\pi}_1$	$\frac{\hat{\pi}_1}{\hat{\psi}_1}$	$p(\hat{\psi}_1 = \hat{\pi}_1)$
Years of Schooling	0.81**	0.35**	0.44	0
Cognitive Ability	3.16**	1.7**	0.54	0.02
CESD Depression Index	0.13**	0.06	0.42	0.25
Birth Weight	3.1**	3.26*	1.05	0.91
Body Mass Index	2.01**	2.34**	1.17	0.59
Height	2.5**	2.51**	1	0.97

* 0.05 ** 0.01 . All models control for sex, age, and the first 10 principal components of individual genotype. All model uses only individuals of European ancestry. Models without individual-fixed effects use a sample of unrelated individuals, whereas the family-fixed effect models use a sample of sibling pairs. The sample of unrelated individuals contains one randomly selected sibling from each sibling pair. All polygenic scores are both standardized within sample to be mean 0 and standard deviation 1. Cognitive ability is measured through the Peabody Picture Vocabulary Test during Wave 1 of Add Health, when respondents were approximately 16 years old. Birth weight is retrospectively reported by respondents' parents during Wave 1 of Add Health. Birth weight is retrospectively reported by respondents' parents during Wave 1 of Add Health. Years of schooling, CESD depression index, body mass index, and height are measured during Wave 4 of Add Health, when respondents were approximately 28 years old. Height is reported in centimeters, birth weight is reported in ounces, and cognitive ability is reported in IQ score points. The CESD depression index is normalized to be mean 0 and standard deviation 1. A traditional regression table is available in Section A12 of the appendix. P-values for the test that $\hat{\psi}_1 = \hat{\pi}_1$ were calculated through bootstrap resampling with replacement simulated with 1000 repetitions.

Figure Legends

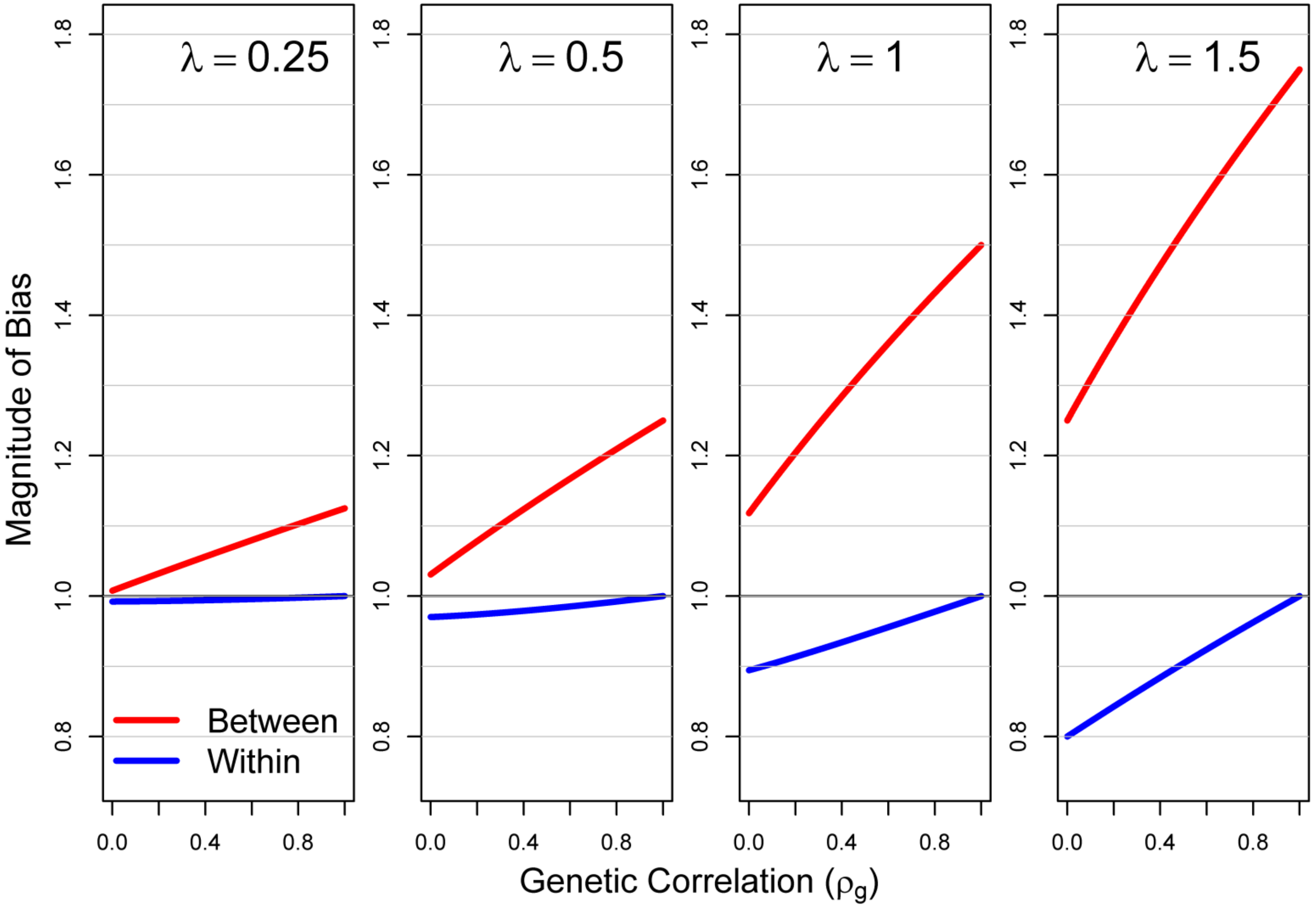
Figure 1

Title: Bias Due to Genetic Nurture in Within and Between-family Regressions using Polygenic Scores

Caption: Grey line at y=1 represents no bias. ρ_g is the direct-nurture genetic correlation and λ is the direct-nurture heritability ratio. Results are derived analytically from a theoretical model.

Figures

Figure 1


λ = 0.25
λ = 0.5
λ = 1
λ = 1.5
Between
Within
Magnitude of Bias
Genetic Correlation (ρ_g)