Social-Science Genomics: Progress, Challenges, and Future Directions

By Daniel J. Benjamin, David Cesarini, Patrick Turley, and Alexander Strudwick Young*

Rapid progress has been made in identifying links between human genetic variation and social and behavioral phenotypes. Applications in mainstream economics are beginning to emerge. This review aims to provide the background needed to bring the interested economist to the frontier of social-science genomics. Our review is structured around a statistical framework that nests many of the key methods, concepts and tools found in the literature. We clarify key assumptions and appropriate interpretations. After critically reviewing several significant applications, we conclude by outlining future advances in genetics that will enable more and improved applications, and we discuss the ethical and communication challenges that arise in this area of research.

* Benjamin: Behavioral Decision Making Group, UCLA Anderson School of Management, and Human Genetics Department, UCLA David Geffen School of Medicine, Los Angeles and NBER (e-mail: daniel.benjamin@gmail.com). Cesarini: Center for Experimental Social Science, New York University, Department of Economics, New York University, NBER and IFN (e-mail: david.cesarini@nyu.edu). Turley: Center for Economic and Social Research and Economics Department, University of Southern California (e-mail: pturley@usc.edu). Young: Human Genetics Department, UCLA David Geffen School of Medicine, Los Angeles (e-mail: alextisyoung@gmail.com). For helpful comments and suggestions, we thank Jonathan Beauchamp, Graham Coop, Molly Przeworski, Anastasia Terskaya, Carl Veller, and Peter M. Visscher and the University of Queensland Statistical Genomics Lab Meeting. We are grateful to Matthew Howell, Giorgia Mezzetti and Moeen Nehzati for research assistance. For financial support, we thank the NIA/NIH (grants R24-AG065184, R01-AG042568, R00-AG062787, and R01-AG081518) and Open Philanthropy.

Over the past 15 years, social-science genomics—a field at the intersection of genetics and social science—has undergone a profound transformation. Prior to the availability of measures of genetic variation, most research treated genetic influences as latent variables and sought to infer their effects by contrasting the resemblance of twins, adoptees, and other kinships **on outcomes of interest** (Goldberger, 2005; Sacerdote, 2011; Cloninger, Rice and Reich, 1979). Advances in genotyping technologies, including an exponential decline in the cost of measuring genetic variation (Wetterstrand, 2023), and concomitant improvements in statistical methods have made entirely new study designs possible. Today, genome-wide association studies (GWASs) allow researchers to directly estimate associations between individual genetic variants and **individuals' outcomes** in large samples. These studies have identified many novel, replicable, genetic associations. In independent samples of genotyped individuals, estimates from a GWAS can be used to aggregate genetic variants across the genome into polygenic indexes (PGIs) that can have substantial predictive power. This methodological revolution, together with the proliferation of datasets incorporating genotypic information, has catalyzed an explosion of research, mostly in medicine and epidemiology but increasingly also in the social sciences, including economics. **Unfortunately, obtaining the large samples needed for GWAS necessitated moving away from the causal research designs of kinship studies toward correlational designs. In our view, the most exciting recent methodological development has been the burgeoning availability of genotyped family samples, which has made it possible to build causal designs into GWAS and PGI studies.**[1]

Much of the recent work in social-science genomics, although constrained to correlation analyses, aims to address causal questions about the role of genetic variation and its interaction with environmental variation in shaping human behavior and socioeconomic outcomes. Examples include: How much do genes influence socioeconomic outcomes? What share of the health-socioeconomic-status gradient is due to genes? Can education buffer individuals with high genetic risk for Alzheimer's disease? Can PGIs be used to identify individuals who might benefit the most from targeted interventions? Do parents offset or amplify genetic differences between siblings? How does assortative

[1]**Different names for this new area of research have been adopted in different social-science disciplines, reflecting their focus on particular applications and methodological approaches. In economics, genoeconomics (Benjamin et al., 2007); in political science, genopolitics (Fowler and Dawes, 2013); and in sociology, sociogenomics or social genomics (Mills and Tropf, 2020; Conley, 2016); however, sociogenomics and social genomics also describe a separate field of research on how social processes affect gene expression, as in Robinson, Grozinger and Whitfield, 2005). Psychology has a longstanding tradition of research on the role of genetics called behavior genetics (for a history, see Loehlin, 2009), and the psychologically oriented research we mention in this review is part of that literature. Although this review reflects our economics perspective and the applications we discuss are economics-oriented, we use the more general term, social-science genomics, which may have originated in Rietveld et al. (2013), to emphasize the commonality of theory, data, and tools—which are what we focus on in this review.**

mating shape genetic and environmental contributions to inequality? The main goal of this review is to equip economists with the theoretical and empirical tools necessary to engage with this burgeoning field**, with a focus on how to incorporate the newly available opportunities for causal inference into the research.**[2]

In addition to bringing interested researchers to the frontier of social-science applications, this paper also aims to fill the need for a textbook treatment of the foundational conceptual, interpretational, and methodological issues in social-science genomics and thereby serve as a resource for economists (and other social scientists) who want to incorporate genetic data into their research. Our review is organized around a statistical framework that we use to establish a common language and formally define key parameters and concepts. The framework clarifies a number of common misunderstandings and provides a useful way to interpret the coefficient estimates from regressions of outcomes on genetic variables. For example, we use the framework to discuss key identifying assumptions underlying various approaches to causal inference, highlight the value of family-based data, and clarify intuitions underlying results. We also use the framework to discuss conceptual pitfalls that arise when interpreting genetic and environmental effects. Finally, we use the framework to describe a number of methods used in social-science applications.

This review is organized as follows. Section I begins with a rudimentary genetics primer. Section II develops our theoretical framework. Although economists will be more interested in applications than in details of genetics research, appropriate interpretation of the applications requires understanding some nuances of how genetics researchers estimate genetic effects; we describe the key issues in Section III. Section IV defines, interprets, and analyzes PGIs, the centerpiece of most applications in recent years. In Section V, we critically review applications of genetic data in economics, as well as some opportunities for future applications. Because the field is advancing so quickly, most of the existing applications are already out of date methodologically; we highlight how future work could improve on them. In Section VI, we outline current trends in genetics research and what they imply about future applications in the social sciences. We conclude by highlighting some of the ethical, policy, and communication challenges that are

[2]In light of the advances over the past few years, we believe the time is ripe for a review paper. **Three** prior review papers in economics journals (Beauchamp et al., 2011; Fletcher, 2011; Benjamin et al., 2012) were published before the first large-scale GWAS of a social-science **outcome** (Rietveld et al., 2013). A third (Dias Pereira et al., 2022) provides an accessible, succinct, and non-technical introduction to a subset of the topics we discuss here. Reviews have also been published in sociology (Freese, 2018; Braudt, 2018; Martschenko, Trejo and Domingue, 2019; Conley, 2016) and psychology (Plomin et al., 2016). Relative to these, we seek to spell out connections to relevant genetic theory more explicitly and to provide a more self-contained and comprehensive treatment of technical details. Although our review is primarily oriented toward economists—for example, most of the applications we highlight are from economics—it is written in the hope that the material will also be of utility to researchers from other disciplines. The theoretical focus of our paper makes it a natural companion to other texts that focus on practical issues that arise in empirical analyses that incorporate genetic data (e.g., a recent textbook by Mills, Barban and Tropf, 2020).

intrinsic to research at the intersection of genetics and social science.

I Genetics Primer

This section provides genetics background relevant to what follows. Some readers may wish to skip the section and refer back to it as needed. Because we aim to provide close to the minimum amount of information needed to fully understand assumptions made elsewhere in the paper, we omit several nuances.[3] For readers interested in additional details, we recommend consulting textbooks in molecular genetics such as Strachan and Read (2018) and population genetics such as Gillespie (2004).

A The Genome and SNPs

The *genome* usually refers to a person's genetic material. Almost every cell in the body contains an exact copy of the entire genome.[4] The human genome has ∼21,000 *genes*. **Each gene is a strip of DNA that provides instructions to the body for how to build a particular protein.** Genes constitute only ∼2% of the genome. A much larger fraction of the genome affects when and how much genes are expressed.

The genome is divided across 23 pairs of *chromosomes*, one sex chromosome pair and 22 non-sex chromosome pairs called autosomes. One chromosome in each pair was inherited from the individual's mother (the maternal chromosome), and the other from the individual's father (the paternal chromosome).

Each chromosome consists of a pair of DNA strands that are bound together. Each strand is composed of a sequence of nucleotide molecules, referred to as bases. There are four bases: guanine (abbreviated G), cytosine (C), thymine (T), and adenine (A). DNA bases always pair with their complementary base on the other strand: C with G, and A with T. Since the information is redundant, one strand is chosen by convention to be the reference strand, and the *base pair* is described by the base on the reference strand.

Each location in the genome can be described by its chromosome and base-pair position. At each position, an individual has one base pair (G, C, T, or A) from the maternal chromosome and one from the paternal chromosome. With rare exceptions, the biological function of the base pair does not depend on whether it was inherited from the mother or father. Thus, the *genotype* at a position—the composition of the genome at that position—can be described by **the pair of parentally inherited** bases (each on its reference strand), such as GC or TT, without reference to which was inherited from which parent.

[3]For example, we ignore mitochondrial DNA, which is technically part of the human genome but resides in mitochondria (outside the cell nucleus) and is inherited exclusively from the mother. We ignore it because mitochondrial DNA only contains 13 of the ∼21,000 human genes and is not generally included in the genotyping data we discuss.

[4]One important exception is germ cells, discussed in Section I.B. We also ignore mutations that cause small differences in DNA across cells.

At the vast majority of **positions**, any two individuals have the same genotype (Nurk et al., 2022). The parts of the genome that vary across individuals are called *genetic variants*. Definitions vary, but according to a typical definition, a *rare variant* is a genetic variant in which 99% or more of individuals have the same version of the genetic variant, and a *common variant* (or a *polymorphism*) is one in which fewer than 99% of individuals have the same version. People's genomes may vary in complex ways from each other; for example, sections may be duplicated, deleted, or inverted. The simplest, and by far most common, type of variation is a single-nucleotide difference, called a *SNP* (*single-nucleotide polymorphism*). **For example, at a SNP that has two possible nucleotides (on the reference strand), A and G, there are three possible genotypes an individual could have: AA, AG, and GG (recall that GA is the same as AG, since which was inherited from which parent does not matter). The nucleotides that can occur at a SNP (A and G in the above example) are called alleles.**

At the vast majority of SNPs, there are only two alleles of non-negligible frequency in the population. Whichever allele is less common in the population is called the *minor allele*. An individual's SNP genotype is often summarized by the minor allele count: 0, 1, or 2. **If an individual's SNP genotype is 0 or 2, then the individual's two alleles are the same, and the individual is referred to as *homozygous* at that SNP. If an individual's SNP genotype is 1, then the individual has one of each of the two possible alleles, and the individual is referred to as *heterozygous* at that SNP.**

SNP data for an individual typically comes as a vector of minor allele counts, with each element corresponding to a measured SNP at a particular locus. Following standard terminology, we will often refer to the minor allele count as the *genotype* and the vector of minor allele counts as the *genotype vector*.

B Genetic Inheritance

For reproduction, individuals produce *germ cells* (sperm in males, eggs in females) via a type of cell division called *meiosis*. Unlike other cells, germ cells contain only one copy of each chromosome. An offspring is conceived when one germ cell from the father and one from the mother fuse. The resulting child then has a chromosome pair, with one chromosome coming from the father and one from the mother.

The single copy of each chromosome in a germ cell is a random mixture of the parent's two copies of that chromosome. There are two distinct stages of randomness. First, within each chromosome pair, the chromosomes cross a random number of times at random loci, and the chromosomes swap their DNA after the crossing points. This process is called crossing over, and the transfer of chunks of DNA is called *recombination*. Second, independently across the 23 chromosome pairs and after recombination, one among each pair is, with equal probability, transmitted to a given germ cell. This process is called *Mendelian segregation*.

These random processes have some important implications for our purposes. Fixing any given SNP, conditional on the parental genotypes, the offspring receives one of each parent's two alleles, with equal probability. For any two SNPs on different chromosome pairs, the transmission of alleles across the two SNPs are independent random processes. Finally, for any two SNPs on the same chromosome pair, the probability that alleles on the same parental chromosome are transmitted to the offspring is higher the closer the two loci are (because it is less likely that crossing over occurred in between the two SNPs). This correlated inheritance of alleles on the same parental chromosome is called *linkage*.

C Linkage Disequilibrium (LD)

Linkage disequilibrium (LD) refers to correlation between the genotypes of genetic variants.[5] Under random mating, the only source of LD is linkage, and therefore no LD is expected between genetic variants on different chromosomes. Within each chromosome, the LD between two variants is generally decreasing with their physical distance. For nearby variants on a chromosome, the LD due to linkage can be very high, often reaching one or nearly one. Regions of the genome that are essentially perfectly correlated with each other in a given population are called *haplotype blocks*, and the different versions of a block effectively form a single genetic variant for that block.

Non-random mating generates LD with different properties. Consider *assortative mating*: individuals who have some characteristic are more likely to mate with other individuals who have that characteristic. Most assortative mating processes that cause spousal resemblance on a **characteristic** will also induce a correlation between spousal genotypes associated with the **characteristic**. One example is height. Since genetic variants associated with height are scattered throughout the entire genome, assortative mating will lead to positive LD between height-associated alleles, including those located on different chromosomes. Another, related form (e.g., Bergstrom, 2013) of **assortative** mating is *population structure*: individuals within a subpopulation—e.g., geographic region, or with a shared ethnicity or language—tend to mate with each other. In that case, alleles that **are** more common within the subpopulation will become correlated, regardless of their **position** in the genome.

[5] In other areas of genetics, LD refers more generally to the statistical association between genotypes, and measures other than correlation are sometimes used. Originally, LD was more specifically related to linkage than it is in modern usage. The concept of LD arose from considering what would happen after repeated recombinations. For example, consider a population where some individuals have an A allele at locus 1 and a T allele at locus 2 and other individuals have a C allele at locus 1 and a G allele at locus 2. In the next generation, recombination between the loci will reduce the association between having an A allele and a T allele. In the limit after many generations, the genotype at locus 1 will become statistically independent of the genotype at locus 2 and remain so thereafter. This equilibrium state is called "linkage equilibrium." LD was meant to refer to deviations from that state (Sved and Hill, 2018).

D Complex Phenotypes

A *trait*, or *phenotype*, is any measurable characteristic, behavior, or outcome of an organism. A phenotype is called *monogenic* if most or all of the variation is controlled by a single gene. A phenotype is called *polygenic*, or *complex*, if it is affected by many genetic variants, not restricted to a single gene. Intermediate cases, where genetic variation is controlled by several genetic variants, also exist, as do hybrid cases. Late-onset Alzheimer's disease is an example of the latter: a single gene, *APOE*, has a relatively large effect, but most of the genetic influence is polygenic (Lambert et al., 2013).

Monogenic traits are featured in standard introductions to genetics. Classic examples dating back to Gregor Mendel's original experiments (Mendel, 1866) include whether a pea is green or yellow, or whether a pea is smooth or wrinkled. Monogenic diseases include phenylketonuria and Huntington's disease. Until roughly 2005 (when genome-wide association studies (GWASs) began to be conducted), progress in identifying specific genetic variants was restricted to monogenic traits, whose inheritance patterns can be traced through family pedigrees.

Most diseases—indeed, most phenotypes—are complex phenotypes. Examples include height and liability to diseases such as schizophrenia and Type 2 diabetes. Much recent progress in medical genetics has been in the domain of complex phenotypes, based on methods such as GWAS. Because virtually all phenotypes of interest to social scientists are complex phenotypes, this paper focuses on theory and methods relevant to them.

E Genomic Data: Sequencing and Genotyping

For research purposes, genetic data are typically obtained from a saliva or blood sample. The two main technologies for measuring DNA are genome sequencing and genotyping arrays.

Sequencing refers to reading segments of DNA sampled from the genome. Sequencing for clinical diagnostics has high coverage of the clinically relevant genetic variants and is usually highly accurate. For research, most human genetic data today instead comes from *SNP arrays*, which measure a pre-specified set of SNPs. The array is chosen to have high coverage of the haplotype blocks and other common genetic variants in a particular population (or across several populations). Thus, the SNPs measured on an array are correlated with, or "tag," the vast majority of variation in the genome that is due to common variants (including common, non-SNP genetic variants). Typical arrays used today measure roughly 1 million SNPs.

Both sequencing and SNP array technologies have experienced sustained, exponential declines in cost over the past few decades. Almost all data used in genome-wide association studies (see Section III.A) have been from SNP arrays because SNP array genotyping has been much less expensive. Today, such genotyping costs roughly $30 per participant.

II Statistical Framework: Genetic Effects

In this section, we lay out a framework for understanding the relationship between genetic variants and complex phenotypes. The framework builds on classic treatments in Fisher (1918) and Falconer (1960). Relative to earlier work, we have clarified the assumptions and interpretation of the framework, especially when and how it can be interpreted causally, by using modern conceptual apparatus such as the potential outcomes notation (e.g., Rubin, 1974).

A The General Framework

Consider a large population of individuals. Each individual's phenotype of interest is denoted y, and each individual's genotype vector is a row vector denoted

$$\mathbf{x} = \left(x_1, x_2, \ldots, x_J\right) \in \mathbb{R}^J,$$

where x_j (before recentering) represents the number of minor alleles at variant $j = 1, 2, ..., J$. Without loss of generality, we recenter y and each x_j so that they all have mean zero **across individuals** in the population.

To define and analyze causal genetic effects, we think of each possible genotype vector as a "treatment" that an individual could be exposed to.[6] The number of possible genotype vectors, however, is enormous (3^J if each variant has three possible genotypes). If we were analyzing a randomized experiment **with one treatment and a control group,** for each individual we would define one potential outcome if assigned to the treatment group and a possibly different potential outcome if assigned to the control group. Here, we define a potential outcome function $y_i(\cdot)$ that maps each possible genotype vector $\mathbf{x}$ to a phenotype value $y_i(\mathbf{x})$. **For a particular genotype vector $\mathbf{x} = \mathbf{x}_i$, we denote i's resulting phenotype**, y_i, as

$$y_i = y_i(\mathbf{x}_i).$$

The causal effect of a (hypothetical) intervention that changes i's genotype vector from $\mathbf{x}$ to $\mathbf{x}^\star$ is $y_i(\mathbf{x}^\star) - y_i(\mathbf{x})$.

We define the *genetic factor*, $G(\mathbf{x}_i) = \mathbb{E}\left[y_i(\mathbf{x}) \mid \mathbf{x} = \mathbf{x}_i\right]$, as the mean potential outcome function in a population conditional on some **genotype vector $\mathbf{x} = \mathbf{x}_i$, where the expectation is taken across the individuals in the population.** Thus, we can express i's potential outcome at any $\mathbf{x} \in \mathbb{R}^J$ as:

$$y_i(\mathbf{x}) = G(\mathbf{x}) + \nu_i(\mathbf{x}),$$

where $\nu_i(\mathbf{x})$ has mean zero **for all $\mathbf{x}$ and captures all factors that cause**

[6] While not universally accepted, this approach aligns with R.A. Fisher's view. In the single-variant case, he distinguished what he called the "average effect"—the causal effect resulting from a hypothetical experiment with a randomized allelic substitution—from the "average excess"—the regression coefficient from a phenotype–genotype regression (Lee and Chow, 2013).

$y_i(\mathbf{x})$ to differ from the mean potential outcome in the population, $G(\mathbf{x})$. Because $G(\mathbf{x})$ is the mean potential outcome across individuals for genotype vector $\mathbf{x}$, $G(\mathbf{x}^\star) - G(\mathbf{x})$ is the average causal effect (across individuals) of changing the genotype vector from $\mathbf{x}$ to $\mathbf{x}^\star$.

B The Additive Genetic Factor

The genetic factor $G(\mathbf{x})$ defined above depends on the distribution of potential outcome functions $y_i(\cdot)$ in the population, but it does *not* depend on the distribution of actual genotype vectors. In contrast, the additive genetic factor, which we will define in this subsection, depends on both the distribution of potential outcome functions *and* the distribution of genotype vectors in the population. We denote by $\mathbf{X}$ the random variable of genotype vectors in the population, and we subscript expectations and variances by $\mathbf{X}$ when they are taken with respect to the distribution of genotype vectors. The symmetric $J \times J$ variance–covariance matrix of the genotype vector,

$$\mathbf{\Sigma} \equiv \mathrm{Var}_{\mathbf{X}}(\mathbf{X}) = \mathbb{E}_{\mathbf{X}}\left[\mathbf{X}\mathbf{X}'\right],$$

is known as the *linkage disequilibrium (LD) matrix* and has typical element $(\mathbf{\Sigma})_{lk} = \mathbb{E}[X_l X_k]$, representing a variance when $l = k$ and a covariance otherwise. We make the standard assumption that the columns of $\mathbf{\Sigma}$ are linearly independent, or equivalently that the matrix is full rank, so that it is invertible.

The *additive genetic factor* is defined as the least-squares projection of the genetic factor $G(\mathbf{x})$ onto $\mathbf{x}$:

$$g(\mathbf{x}) = \mathbf{x}\boldsymbol{\beta}, \quad \boldsymbol{\beta} \equiv \mathrm{argmin}_{\mathbf{b}}\, \mathbb{E}_{\mathbf{X}}\Big(G(\mathbf{X}) - \mathbf{X}\mathbf{b}\Big)^2, \quad \boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_J)'.$$

From standard arguments (e.g. Goldberger, 1991), given the distribution of genotype vectors in the population, $g(\mathbf{x})$ is the best (in the sense of mean squared error) linear approximation to $G(\mathbf{x})$. An important subtlety is that the distribution of $\mathbf{X}$ will differ in populations with different LD patterns (such as populations of African rather than European genetic ancestries). Therefore, such populations will have different $\boldsymbol{\beta}$ vectors even if their $G(\mathbf{x})$'s are identical. **In a randomly mating population,** when a large number of variants have non-zero effects and no variants exhibit effect sizes that are outliers, the additive genetic factor will be approximately normally distributed.

The deviation of the genetic factor $G(\cdot)$ from the best *linear* approximation $g(\cdot)$ in a population is known as the *non-additive genetic factor*, $N(\mathbf{x}) \equiv G(\mathbf{x}) - g(\mathbf{x})$. The deviations from linearity are often decomposed into two terms: dominance and epistasis. *Dominance* refers to non-linearity in the effects of a genetic variant **(for example, changing the minor allele count from 0 to 1 has a larger effect on the phenotype than changing it from 1 to 2).** *Epistasis* (or

gene-gene interaction effects) refers to interactions between the genotypes of two or more genetic variants.

In most of what follows, our focus is on the *additive model*:

$$y_i(\mathbf{x}) = \mathbf{x}\boldsymbol{\beta} + \epsilon_i(\mathbf{x}), \tag{1}$$

where the residual is $\epsilon_i(\mathbf{x}) = N(\mathbf{x}) + \nu_i(\mathbf{x})$. By the properties of linear regression, $\mathbb{E}_{\mathbf{X}}[N(\mathbf{X})] = 0$ and $N(\mathbf{X})$ **is uncorrelated with** each X_j (and any linear combination of $X_1, X_2, ..., X_J$). However, $\epsilon_i(\mathbf{X})$ may nonetheless be correlated with $\mathbf{X}$ if $\nu_i(\mathbf{X})$ is.

The elements of the vector $\boldsymbol{\beta}$ can be interpreted as average causal effects: $G(\cdot)$ **represents average causal effects across individuals; and as the best linear approximation of** $G(\cdot)$**,** $g(\cdot)$ **represents the causal effects** averaged over non-linearities. This interpretation remains valid if the true effects vary across individuals, e.g., due to gene-by-environment interactions. Moreover, deviations from linearity are captured by $N(\mathbf{X})$, which can be treated as error because it has mean zero and is orthogonal to $\mathbf{X}$. Thus, the average-causal-effects interpretation does not hinge on restrictive assumptions about the functional form of the potential-outcome functions.

That said, the usefulness of knowing $\boldsymbol{\beta}$ is presumably greater when the approximation of $G(\mathbf{X})$ by $g(\mathbf{X})$ is more accurate. Theoretical arguments in statistical genetics imply that, for complex phenotypes **(unlike for monogenic traits), the approximation should often be quite good (Hill, Goddard and Visscher, 2008). To be more precise, the difference between** $G(\mathbf{X})$ **and** $g(\mathbf{X})$**—i.e., the non-additive genetic factor, comprised of dominance and epistasis—is anticipated to explain much less of the variance in the phenotype than the additive genetic factor** $g(\mathbf{X})$. The available empirical evidence confirms **this prediction**. For example, one study of 50 distinct traits conducted in UK Biobank found that the proportion of variance **in the phenotype** explained by dominance deviations is consistently smaller than 1%, compared to an average of 21.9% for the additive genetic factor (Pazokitoroudi et al., 2021). Empirical research on epistatic interactions has similarly failed to provide compelling evidence for substantial epistatic variance, but due to the combinatorial challenges, estimation is limited by statistical power and computational complexity, leading to wider confidence intervals (Hivert et al., 2021).

In summary, the additive genetic model's appeal stems from two main sources: (i) its tractability, requiring only one parameter per SNP, β_j; and (ii) the additive genetic factor $g(\mathbf{x})$ is often a close approximation to the genetic factor $G(\mathbf{x})$.

C Unmeasured Variants and the Additive SNP Factor

In most empirical analyses of genetic data, only a strict subset of the J elements of $\mathbf{x}$ are observed, where virtually always all $K < J$ **elements of the subset** are SNPs. For such cases, the additive genetic factor $g(\mathbf{x})$ is sometimes approximated

by what we call the *additive SNP factor*, $\tilde{g}(\mathbf{x})$, which is defined analogously but for the observed genotypes. The observed genotype vector is the $1 \times K$ subvector $\tilde{\mathbf{x}} = \left(\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_K\right)$, and $\tilde{\mathbf{X}} = \left(\tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_K\right)$ **is the random variable consisting of the elements of X that correspond to observed genotypes**. The *additive SNP factor* is then given by:

$$\tilde{g}(\tilde{\mathbf{x}}) = \tilde{\mathbf{x}}\tilde{\boldsymbol{\beta}}, \quad \tilde{\boldsymbol{\beta}} \equiv \text{argmin}_{\mathbf{b}}\, \mathbb{E}_{\mathbf{X}}\Big(G(\mathbf{X}) - \tilde{\mathbf{X}}\mathbf{b}\Big)^2, \quad \tilde{\boldsymbol{\beta}} = (\tilde{\beta}_1, \tilde{\beta}_2, \ldots, \tilde{\beta}_K).$$

The definition is identical to that of the additive genetic factor, the *only* difference being that the least-squares projection is onto $\tilde{\mathbf{x}}$ instead of $\mathbf{x}$. **Because it is a least-squares projection of causal genetic effects, variation in the additive SNP factor $\tilde{g}(\mathbf{x})$ across individuals captures variation in causal genetic effects. However, because the projection is onto a subset of the causal genetic variatns**, the analog of Equation (1) cannot be expressed in terms of potential outcomes, $y_i(\mathbf{x})$. Instead, we write the additive SNP model as:

$$y_i = \tilde{\mathbf{x}}_i\tilde{\boldsymbol{\beta}} + \tilde{\epsilon}_i. \tag{2}$$

Intuitively, the causal interpretation of $\boldsymbol{\beta}$ does not extend to $\tilde{\boldsymbol{\beta}}$ because there is omitted-variable bias: when some variants are omitted **from the least-squares projection, the value of any $\tilde{\beta}_k \in \tilde{\boldsymbol{\beta}}$ will typically reflect some mix of the average causal effect of SNP k itself and the average causal effects of unmeasured genetic variants correlated with $\tilde{x}_k$. Therefore, the proportion of variance in the unmeasured genetic variants that is captured by (linear combinations of) the K observed SNPs depends partly on the LD matrix $\boldsymbol{\Sigma}$**.

D Genetic Factors, Heritability and Genetic Correlation

Heritability is often described as the proportion of variation **in a phenotype** "attributable" to genes, whereas the term *genetic correlation* is used to characterize the degree of "genetic overlap" between two phenotypes. While informal definitions such as these can be helpful, they ignore distinctions between different notions of heritability and genetic correlation. Understanding the distinctions matters for at least two related reasons. First, the different notions can lead to dramatically different parameter values, especially for heritability. For example, twin studies often estimate the heritability of adult height to be around 90% (e.g. Silventoinen et al., 2003), whereas studies with genome-wide data often estimate it to be around 45% (Yengo et al., 2022). These numbers give divergent impressions of how "genetic" height is, but mainly the difference arises because the studies are estimating different parameters. Second, different notions are relevant depending on the research question.

There are three main notions of a phenotype's heritability. Each is the ratio of

the variance in a factor to the variance **in the phenotype:**

$$\frac{\text{Var}(F(\mathbf{X}))}{\text{Var}(y)} = \begin{cases} H^2, & \text{if } F(\mathbf{X}) = G(\mathbf{X}) \quad \text{(broad heritability)} \\ h^2, & \text{if } F(\mathbf{X}) = g(\mathbf{X}) \quad \text{(narrow heritability)} \\ \tilde{h}^2, & \text{if } F(\mathbf{X}) = \tilde{g}(\tilde{\mathbf{X}}) \quad \text{(SNP heritability)} \end{cases}$$

where the variances are taken across individuals in the population. The three measures have several features in common. Because variation across individuals in a factor represents variation in causal genetic effects, each measure of heritability captures some notion of how much variation in the phenotype is *caused* by genetic variation. Specifically, each is interpretable as the R^2 that would be obtained if it were possible to run a population regression of the phenotype onto a factor (this regression is infeasible because the factor is unobserved). Since allele frequencies, causal genetic effects, and environmental conditions vary across populations, and since each of these could affect the R^2 of such a regression, heritabilities should generally be expected to vary across populations.

To understand how the three notions relate ordinally, observe that $G(\mathbf{x})$, $g(\mathbf{x})$ and $\tilde{g}(\tilde{\mathbf{x}})$ can be interpreted as solutions to the problem of finding the best predictor of $y_i(\mathbf{x})$ **(in the sense of minimizing mean squared error)** with additional constraints successively imposed. The unconstrained solution is $G(\mathbf{x}) = \mathbb{E}[y_i(\mathbf{x})|\mathbf{x}]$. The step $G(\mathbf{x}) \to g(\mathbf{x})$ comes from imposing the constraint of a linear function $g(\mathbf{x}) = \mathbf{x}\boldsymbol{\beta}$. In the step $g(\mathbf{x}) \to \tilde{g}(\tilde{\mathbf{x}})$, the linear approximation is further restricted to the K observed SNPs only: $\tilde{\mathbf{x}}\tilde{\boldsymbol{\beta}}$. It follows that

$$H^2 \geq h^2 \geq \tilde{h}^2,$$

where the first inequality is strict except in the special case where dominance and epistasis are both absent. Under our full-rank assumption, the second inequality is strict whenever at least one of the unmeasured variants has a nonzero effect on the phenotype. **The 90% estimate for height quoted above is an estimate of h^2 (which, for height, is believed to be very close to H^2), whereas the 45% estimate quoted above is an estimate of $\tilde{h}^2$.**

Broad-sense heritability, H^2, is the closest to an overall assessment of genetic influence on variation (albeit subject to caveats below). Narrow-sense heritability, h^2, is the key parameter for quantifying the response to selection pressures, both natural and artificial (Visscher, Hill and Wray, 2008). For example, it is used in plant and animal breeding to predict how selective breeding will change the trait distribution across generations. Outside these settings, to the extent that h^2 remains stable over time, it can provide an upper bound on the accuracy of *future* genetic predictors within the same population (that are built as

linear combinations of genetic variants, including those not measured by current technologies). This upper bound can be used, **for example,** in cost-benefit analyses of precision-medicine initiatives or in examinations of the conditions under which advances in genetics research could lead insurance markets to unravel by disrupting risk pooling (see Section V.C).

SNP heritability, $\tilde{h}^2$, provides an upper bound on the accuracy of *current* genetic predictors within the same population that are based on (linear combinations of) the K observed SNPs. Knowing this upper bound can be useful in social-science applications, as we discuss in Section IV.E. When comparing SNP heritabilities, a complication is that the parameter value will generally depend on which SNPs are observed. In practice, differences in genotype array, **strategy for imputing unobserved genetic variants, and quality-control filters (i.e., dropping some measured SNPs)** generate such a barrier to comparability, a challenge we return to in our discussion of polygenic indexes (Section IV).

We now turn to the two main notions of the genetic correlation between a pair of phenotypes, m and n (Okbay et al., 2016; Border et al., 2022*b*). The *factor-based* genetic correlation is the *within-person* correlation between the additive genetic factors for m and n: $\rho(\mathbf{x}\boldsymbol{\beta}_m, \mathbf{x}\boldsymbol{\beta}_n) = \frac{\boldsymbol{\beta}'_m \boldsymbol{\Sigma} \boldsymbol{\beta}_n}{\sqrt{\boldsymbol{\beta}'_m \boldsymbol{\Sigma} \boldsymbol{\beta}_m \boldsymbol{\beta}'_n \boldsymbol{\Sigma} \boldsymbol{\beta}_n}}$. The *coefficient-based* genetic correlation is the (uncentered) correlation of the causal effects on phenotypes m and n, the correlation being calculated across *all* $j = 1, \ldots, J$ genetic variants: $\rho(\beta_{m,j}, \beta_{n,j}) = \frac{\boldsymbol{\beta}'_m \boldsymbol{\beta}_n}{\sqrt{\boldsymbol{\beta}'_m \boldsymbol{\beta}_m \boldsymbol{\beta}'_n \boldsymbol{\beta}_n}}$. **(The effect size of each SNP is fixed in a population; this correlation can be thought of as how correlated the effect sizes are when you choose a SNP uniformly at random from the genome.)** The factor-based genetic correlation addresses the question: to what extent do individuals who have a higher genetically-influenced propensity for one phenotype also have a higher genetically-influenced propensity for another phenotype? Factor correlation is also relevant when considering whether polygenic indexes are likely to be correlated. In contrast, the coefficient-based genetic correlation addresses the question: to what extent do genetic variants have overlapping causal effects? This correlation is more relevant when the goal is to elucidate causal mechanisms.

In practice, however, the distinction between these two notions is rarely acknowledged in the literature. In part, this may be because data constraints only allow researchers to use methods that estimate one of these. It may also be because empirical estimates are often quite close numerically (Lee et al., 2018; Bulik-Sullivan et al., 2015*a*). **To understand the numerical relationship between the two parameters, note that the formulas only differ by the inclusion of the weight matrix $\boldsymbol{\Sigma}$ in the factor-based correlation. Consequently**, the two parameters differ if the LD between pairs of SNPs is systematically related to whether the SNP effects have concordant signs. For example, if SNP genotypes that increase phenotype m are positively correlated with SNP genotypes that increase phenotype n, then the factor-based correlation (which depends on LD)

will be larger than the coefficient-based correlation (which is independent of LD). Such a relationship can arise for several reasons, such as if mating pairs sort on phenotypes n and m (Border et al., 2022*a*).

Genetic correlations may be of interest to economists **for several reasons. For example,** they can provide new evidence for evaluating theories. For example, Karlsson Linnér et al. (2019) find that coefficient-based genetic correlations between a survey-based measure of general risk tolerance and several risky behaviors across distinct domains exceed the corresponding phenotypic correlations **(i.e., correlation between individuals' phenotype values),** even after adjusting for measurement error that attenuates the phenotypic correlations. The relatively low phenotypic correlations had been interpreted as evidence against the existence of a domain-general risk tolerance parameter (Weber, Blais and Betz, 2002; Hanoch, Johnson and Wilke, 2006). **However, the findings of Karlsson Linnér et al. (2019) mean that the genetic effects on risky behavior are relatively strongly correlated across domains, implying that the lower phenotypic correlations are due to the non-genetic effects (e.g., friend networks that influence propensity to initiate smoking and childhood experiences that influence adventurousness) being weakly correlated. That is, the findings are consistent with a model in which an unobserved, genetically influenced, domain-general tolerance parameter accounts for some of the correlations between distinct risky behaviors (Einav et al., 2016; Frey et al., 2017), but relatively uncorrelated domain-specific non-genetic effects on behavior cause the risky behaviors themselves to have weak correlatives.**

As another example of why genetic correlations may be of interest, the genetic correlation calculated for the *same* phenotype across two different populations can provide evidence on differences in mechanisms across the populations. For example, Martin et al. (2021) estimate the coefficient-based genetic correlation between men and women for 16 different behavioral and psychiatric phenotypes. They find that the genetic correlation is 0.81 (SE = 0.04) for risk-taking behavior and 0.92 (SE = 0.02) for educational attainment. That these correlations are smaller than one implies that the relative importance of the genetic pathways that influence these two phenotypes differs for men and women.

While we have highlighted examples of empirical analyses involving genetic correlations that we believe are plausible, we caution that these examples—like virtually all estimates of genetic correlation to date—are based on study designs with not-fully-compelling identification strategies. As with other research we discuss in this paper, we anticipate that future work will use stronger identification strategies made possible by genotyped family data.

E Interpretational Issues

We now use our statistical framework to clarify several persistent misunderstandings. To isolate these issues, we purposefully set aside empirical challenges such as measurement problems and imperfect identification strategies in order to focus solely on conceptual issues. Our treatment closely follows Jencks (1980), Jencks and Brown (1977), and Goldberger (1979).

In our framework, we imagine a hypothetical experiment in which alleles are randomly assigned at conception. A genetic effect is defined by the difference in average outcomes between groups assigned to different genotypes. Crucially, this difference is a reduced-form parameter: it captures all causal pathways, irrespective of their complexity and the nature of the mediating variables.

Jencks (1980) proposed a taxonomy of three broad pathways through which genes may influence outcomes.[7] We illustrate each using a concrete example where the outcome is body mass index (BMI). First, genes could affect the environments people select into. For example, genetic influences on food preferences could lead individuals to choose different diets, with downstream effects on BMI. Second, genes could evoke environmental reactions. For example, genes that influence physical appearance may affect how a person is treated in various social contexts and ultimately also their BMI. Third, genes could influence outcomes through "physical" biological mechanisms. For example, genes involved in insulin signaling might affect BMI by altering glucose metabolism and fat storage.

Many misunderstandings arise because the standard terminology carries connotations in everyday language that diverge from the technical definitions. For instance, the term "genetic effect" evokes a physical mechanism like those described in biology textbooks for **monogenic** disorders. Jencks (1980) termed the assumption that genetic effects must operate through physical pathways the "genetics = physical" fallacy.

This fallacy leads naturally to a false dichotomy between genetic and environmental causes. A thought experiment proposed by Jencks and Brown (1977) and further developed by Jencks (1980) provides a concrete example. Consider how the effect of being endowed with two X chromosomes at conception (i.e., of being biologically female) has impacted educational outcomes at different points in time. A century ago, the effect of possessing two X chromosomes on educational outcomes was large and negative. Over subsequent decades, male–female disparities in education narrowed in many countries, **in some countries** reversing. A

[7] Jencks's typology is similar to the taxonomy proposed in a well-known paper by Plomin, DeFries and Loehlin (1977), but the focus and conclusions of the two papers are quite different. The first two mechanisms of environmental mediation in Jencks's taxonomy would generate what Plomin, DeFries and Loehlin (1977) informally refer to as "active" and "reactive gene-environment correlations." In Jencks's framework, a correlation between the genetic factor and the exogenous environment corresponds to what Plomin, DeFries and Loehlin (1977) call a "passive gene–environment correlation." Jencks does not adopt formal potential outcomes notation, but the key feature that distinguishes his analysis is that it is explicitly and consistently grounded in counterfactual reasoning.

major mechanism was the gradual erosion of formal and informal barriers that historically constrained women's educational opportunities. This environmental mechanism was surely far more important than some change in the "physical" biological effect of having two X chromosomes.

The false dichotomy between genetic and environmental causes is stubbornly persistent, even among prominent researchers. For example, the abstract of a high-profile paper on the genetics of obesity states: "Although often attributed to unhealthy lifestyle choices or environmental factors, obesity is known to be heritable" (Khera et al., 2019). Yet as already noted, a gene may exert its effect on BMI entirely or in part through effects on "lifestyle choices or environmental factors" such as exercise or diet.

Whenever genes exert some of their effects through environmental variables, it is wrong to interpret one minus heritability as the share of phenotypic variance explained by the environment. Jencks (1980) therefore proposed distinguishing between two types of environmental effects: those whose "ultimate" cause is genes—the *endogenous environment*—and those whose "ultimate" cause lives in the residual—the *exogenous environment*. For example, Sanz-de Galdeano and Terskaya (2025) find that children's genotypes evoke parental investments, implying that some portion of parental investments is endogenous and therefore part of the genetic factor. There is an additional complexity: the residual may *not* be entirely environmental. All three residual terms in our framework—$\nu_i(\mathbf{x})$ in the general model, $\epsilon_i(\mathbf{x})$ in the additive model, and $\tilde{\epsilon}_i$ in the additive SNP model—can encompass sources of variation that are not unambiguously "environmental," including gene-environment interactions (in all three cases $\nu_i(\mathbf{x})$, $\epsilon_i(\mathbf{x})$, and $\tilde{\epsilon}_i$), non-additive genetic variance (in the cases of $\epsilon_i(\mathbf{x})$ and $\tilde{\epsilon}_i$), and effects from unmeasured variants (in the case of $\tilde{\epsilon}_i$). Therefore, one minus the heritability is neither an upper nor a lower bound for the contribution of the environment to variance **in the phenotype.**

Another common misconception is that heritability is informative about the effectiveness of policy interventions. The source of this error is thinking that because the genotype vector $\mathbf{x}$ is fixed, its effects are fixed. However, many mechanisms may be modifiable. For example, biological mechanisms *need not* be immutable. Indeed, perhaps the primary motivation for studying genetic associations with complex diseases is to guide drug discovery efforts, through the identification of more promising therapeutic targets. By now, there is overwhelming evidence that such research has helped improve drug targeting (Fang et al., 2019; Wang et al., 2024). In short, broad categorical distinctions of causes—such as biological vs. social or genetic vs. environmental—are of far more limited use for assessing modifiability or the scope for intervention than is commonly recognized.

Moreover, suppose it could somehow be established that all genetic effects on a phenotype operate through pathways that cannot be modified. Even then, high heritability would *not* imply that there is no scope for policy to be effective. Goldberger noted (1979, p. 344): "The policy-relevant effect of an explanatory

variable is properly measured by its regression slope, not by its contribution to R^2..." In his example, supplying people with eyeglasses can improve vision irrespective of what fraction of the variation is caused by genes. Another example is height. The broad-sense heritability of height is estimated as high as 90%, and yet average height has increased substantially over time as nutritional conditions improved (Visscher, 2008).

In highlighting these conceptual pitfalls, our aim is not to sow doubt about the value of research in this area. On the contrary, the complexities point to fertile areas for empirical and theoretical work in the social sciences. For example, there is substantial interest in understanding how and why heritability varies across time and space (e.g. Branigan, McCallum and Freese, 2013; Rimfeld et al., 2018), how genetic factors moderate environmental effects (Barcellos, Carvalho and Turley, 2018; Miao et al., 2022*a*; Basu et al., 2025), and how children's genes can trigger parental reactions that blur the boundaries between "genetic" and "environmental" effects (Sanz-de Galdeano and Terskaya, 2025). Indeed, we believe an important long-run goal for the field is to advance our knowledge of the structural relationships between the elements of $\mathbf{x}$ and important outcomes.

The complexities do, however, underline the importance of defining terms such as "genetic effect" and "heritability" clearly and explicitly and reminding readers about their nuances. When there is potential for confusion, the term "environment" should be prefixed by exogenous or endogenous. Misunderstandings should be preempted by steering clear of terminology that invites intuitive but incorrect interpretations, or at least explicitly pointing out the imperfect alignment. For example, in economic applications, it is common to refer to the genetic factor (or a polygenic index) as a "genetic endowment." We view this label as imprecise because the genetic factor includes not only the vector of genotypes $\mathbf{x}$, which may reasonably be considered part of the endowment in a human capital model, but also elements like expected endogenous investments, which are a distinct component of the model. In our view, it is better to use the less familiar but more precise term "genetic factor" and carefully explain the link to the theoretical framework that motivates the application.

F Parental, Sibling, and Other Interpersonal Genetic Effects

Many questions in the social sciences are fundamentally about how one person's behaviors or characteristics influence others. For example, how do parental behaviors influence children's outcomes? How does one's college roommate affect academic performance? Under what circumstances does a person's smoking behaviors influence the smoking behaviors of others? A central obstacle to empirically addressing such questions is the reflection problem (Manski, 1993), the simultaneity of mutual influences. Genetic data can help in this domain. Specifically, up to this point, we have analyzed how the elements of $\mathbf{x}_i$ impact i's own

outcome, y_i, which we call *self genetic effects*.[8] In principle, the same framework can be used to define and analyze the causal effect of $\mathbf{x}_i$ on the outcomes of others, which we call *interpersonal genetic effects*. Because a person's genotype vector is fixed at conception, interpersonal genetic effects sidestep the reflection problem.

Interpersonal genetic effects could include *sibling*, *parental, grandparental,* or *friend* (e.g., Sotoudeh, Harris and Conley, 2019) *genetic effects*. We name the genetic effects such that the named person's genes are affecting the focal individual. For example, a grandparental genetic effect is the average effect of a change in a grandparent's genotype on the phenotype of their grandchild. Below we discuss two types of interpersonal genetic effects, sibling and parental, to highlight some nuances of interpretation that can arise in studies of interpersonal genetic effects.

With sibling genetic effects, two primary subtleties arise. First, the relevant population for which sibling genetic effects are well-defined is only the population of individuals who have siblings, whereas self genetic effects are well-defined for the full population. Second, more than one causal parameter of interest may exist. For example, in families with three or more children, one parameter corresponds to the experiment of changing a random sibling's genotype holding other siblings' genotypes constant, while another parameter corresponds to the experiment of changing all siblings' genotypes.

Parental genetic effects may be of particular interest to economists in light of the extensive literatures on parental investments and intergenerational mobility. Conceptually, however, they are substantially more complicated than sibling genetic effects. Most fundamentally, changing a parent's genotype at conception may affect the parent's fertility, mate choice, or timing of children, meaning the focal child may never exist. For parental genetic effects to be well-defined, such effects must be assumed to be negligible. The next issue is that part of the effect of changing a parent's genotype is that the altered genotype may be passed on to the offspring, which could lead to a self genetic effect in the offspring. Because this pathway is a mechanical function of the self genetic effect for the child, parental genetic effects are generally *defined* to be the part of the effect of changing a parent's genotype that does not operate through genetic transmission to the child (for formal treatments, see Shen and Feldman, 2020; Young, 2023; and Veller and Coop, 2024). Under this definition, parental genetic effects are understood to capture the self genetic effects on the parent's phenotypes (e.g., their income and behaviors) that could affect the phenotype of the offspring, and such pathways are indeed relevant to understanding causal effects of parental characteristics. However, there is one more subtlety that is generally underappreciated: if the parent has multiple children, the altered genotype could also be inherited by any of the child's siblings, which could produce sibling genetic effects on the child's

[8] In the literature, what we call self genetic effects are sometimes called "direct genetic" effects, and what we call interpersonal genetic effects are variously called "indirect genetic," "associative," or "genetic nurture" effects. For a review of the genetics literature, almost entirely focused on non-human examples, see chapter 22 in Walsh and Lynch (2018).

phenotype (Young et al., 2022). Therefore, parental genetic effects as generally defined also partially include sibling genetic effects; isolating the part that operates through parental characteristics would require subtracting out the sibling part.

In summary, understanding the magnitudes and mechanisms of interpersonal genetic effects offers the promise of becoming a broadly useful approach to learning about how individuals are affected by the behaviors and environments generated by people around them. Moreover, similar identification strategies that can be used to estimate self genetic effects can be used to credibly estimate interpersonal genetic effects (see Section V.E). However, as with self genetic effects, more work—which usually must rely on weaker identification strategies—is then required to understand the pathways through which the effects operate.

III Estimation of Genetic Effects

Although estimation of genetic effects *per se* is of less interest to economists than social-science applications that use genetic data, we believe **the challenges of estimating genetic effects are** nonetheless valuable to understand because the applications rely on genetic-effect estimates. Consequently, the appropriate interpretation of applications generally requires understanding the limitations of these estimates.

The causal effects of interest are given by the vector $\boldsymbol{\beta}$, as defined in Section II.B. Suppose first that we have a very large sample with all J genetic variants measured and that the genotype matrix has linearly independent columns. Even in this idealized scenario, we face the usual "fundamental problem of causal inference": only one potential outcome is observed per individual. Consequently, we cannot operationalize Equation (1). If we estimate the analog of Equation (1) in terms of observables,

$$y_i = \mathbf{x}_i\boldsymbol{\beta} + \epsilon_i, \tag{3}$$

the ordinary least squares estimator of $\boldsymbol{\beta}$ will be biased if $\mathrm{Cov}\left(\mathbf{x}_i, \epsilon_i\right) \neq 0$. The standard solution is find to some vector of controls $\mathbf{z}_i$ such that the genotype vector is as good as randomly assigned conditional on $\mathbf{z}_i$. We dedicate most of this section to discussing some alternative choices of $\mathbf{z}_i$. Traditionally and still today, the most common $\mathbf{z}_i$ is the vector of top principal components from the LD matrix. As we will discuss, recently emerging choices of $\mathbf{z}_i$—namely, parental genotypes and family fixed effects when the sample is restricted to siblings—provide better identification but require family data.

In practice, two additional complications arise. First, only $K < J$ SNPs are observed. Consequently, the estimated **effects** of the K SNPs include omitted-variables bias from the unmeasured genetic variants. Second, the full-rank condition fails; hence estimation of Equation (3) is impossible even after restricting

to the observed SNPs. This short-rank problem arises because the available sample size N of individuals is typically smaller than K and because some SNPs are perfectly correlated with each other. The short-rank problem motivates the standard study design for genetic-effect estimation, the *genome-wide association study (GWAS)*, to which we now turn.

A Genome-Wide Association Studies (GWAS)

A GWAS proceeds iteratively, each time regressing the outcome on one **SNP** (controlling for $\mathbf{z}_i$). More precisely, for **SNP** j, the estimating equation for a GWAS is:

$$y = x_j \beta_j^{GWAS} + \mathbf{z}\gamma_j + \epsilon_j, \tag{4}$$

where we drop the i subscripts to make the notation more compact and where x_j is the person's genotype at SNP j. Because each regression only contains one SNP and a handful of controls, this strategy solves the short-rank problem. However, this one-at-a-time approach introduces additional omitted-variables bias from the other measured SNPs.

The primary output of the K separate regressions are *GWAS summary statistics*:

$$\hat{\boldsymbol{\beta}}^{\text{GWAS}} = \left(\hat{\beta}_1^{\text{GWAS}}, \hat{\beta}_2^{\text{GWAS}}, ..., \hat{\beta}_K^{\text{GWAS}}\right)',$$

together with their standard errors or p-values. In a GWAS, the conventional p-value threshold for statistical significance is 5×10^{-8}, which is called the *genome-wide significance threshold*.[9] The stringent significance threshold, combined with the very small fraction of variance explained by individual SNPs for polygenic phenotypes, means that GWAS sample sizes have had to be large to have adequate power. For example, the largest SNP associations with body mass index (BMI) have an R^2 of roughly 0.003 (Locke et al., 2015); and those with educational attainment, roughly 0.0002 (Okbay et al., 2016). To attain 80% power to detect these effects at the genome-wide significance threshold requires sample sizes of ∼13,000 and ∼200,000 individuals, respectively.

Absent omitted-variables bias from environmental confounding (i.e., assuming the controls $\mathbf{z}$ are sufficient for **credible** causal inference), β_j^{GWAS} can be expressed

[9]This threshold can be understood as the Bonferroni-corrected 0.05 threshold, given that there are roughly 1 million independent statistical tests **in** a GWAS, after accounting for the LD between the >1 million measured and imputed SNPs (Panagiotou and Ioannidis, 2012). Experience indicates that, to date, this threshold has kept the rate of false positives low in GWAS (e.g. Visscher et al., 2012). However, as genotyping technology improves and captures rarer SNPs (which necessarily have weaker LD with other SNPs), GWASs involve more than 1 million independent statistical tests in European-genetic-ancestry **populations** (Wu et al., 2017). Moreover, even with current genotyping technology, there are many more than 1 million independent statistical tests in samples where LD is on average weaker, such as African-genetic-ancestry **populations**. In these cases, lower p-value thresholds will be needed to keep the rate of false positives as low as it has been.

as a function of the elements of the vector $\boldsymbol{\beta}$ from Equation (3):

$$\beta_j^{\text{GWAS}} = \sum_{k=1}^{J} \frac{r_{jk\perp\mathbf{z}}}{r_{jj\perp\mathbf{z}}} \beta_j, \tag{5}$$

where $r_{jk\perp\mathbf{z}}$ is the covariance between the residual genotype of SNP j and the residual genotype of genetic variant k, after residualizing both on $\mathbf{z}$. Since the parameter vector of interest is usually $\boldsymbol{\beta}$, virtually all analyses of GWAS summary statistics use information about the LD structure of the population to transform the GWAS summary statistics into something closer to **an estimate of $\boldsymbol{\beta}$**, typically using some sparse approximation to the true LD matrix, e.g., a simplified matrix where all pairwise correlations between loci whose physical distance exceeds some threshold are set to zero (e.g., Lloyd-Jones et al., 2019).

B Common Sources of Environmental Confounding

Whenever allele frequencies correlate with exogenous environmental factors, another omitted-variables bias arises. Such bias is called *gene-environment correlation*. Two distinct sources of such correlations are widely recognized in the literature.

One classic source is population stratification, which arises when individuals who share genetic ancestry—and thus tend to have similar genotypes—also share environmental exposures that affect the outcome. A well-known hypothetical example, discussed by Lander and Schork (1994) and Hamer and Sirota (2000), involves a poorly designed GWAS of chopstick use, whose "cases" are sampled from an Asian-ancestry population, whereas "controls" are sampled from a different population. Such a study might falsely attribute **causal genetic effects on chopsticks use** to any SNP with systematic allele-frequency differences between the two groups. Such inferences are of course incorrect because ancestry is correlated with a cultural practice that the analysis does not control for, inducing spurious associations between non-causal SNPs and the outcome.

A second potential source of bias arises from interpersonal genetic effects from relatives (Section II.F). For instance, suppose parental nurturing behavior is genetically influenced, and suppose parental nurturing affects offspring cognitive skills. Because an allele that causes nurturing will sometime be inherited by offspring, there will be an association between that allele and cognitive skills in the offspring, even though the offspring's allele does not have a causal effect on the offspring's cognitive skills. A GWAS of cognitive skills in the offspring may find an association with a SNP that is in LD with that allele and falsely attribute a self causal effect.

In the following three subsections, we discuss three choices of controls $\mathbf{z}$ to include in a GWAS that have been, or are beginning to be, used to mitigate such confounding. We evaluate each approach's underlying assumptions, practical fea-

sibility, and relative credibility **for causal inference**.

C Population-Based GWAS with Principal Components

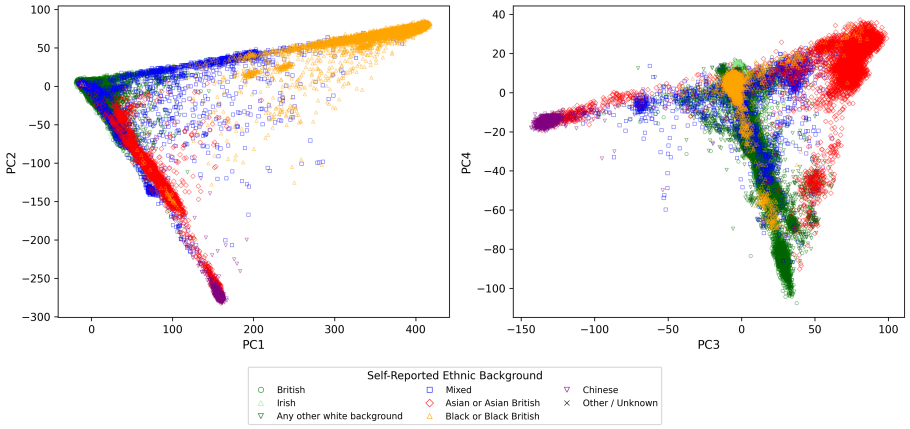
Most human genetic-association studies published to date have been conducted in samples of approximately unrelated individuals with a shared, continental ancestry. Recently, such studies have begun to be referred to as *population-based*, to distinguish them from studies leveraging within-family variation.

In a population-based GWAS, the most common control strategy is to include in $\mathbf{z}$ several genetic principal components (PCs) (in addition to, typically, flexible controls for age and year of birth). These can be inferred from a sample variance-covariance matrix whose $(k, l)^{\text{th}}$ element is the genetic relatedness between individuals k and l, estimated using SNP array data. When PCs are ranked in descending order by eigenvalue, the first PC will capture the greatest proportion of genetic variation in the sample, the second PC will capture the greatest remaining variation orthogonal to the first, and so on.

The top PCs typically capture information about geographic and ethnic ancestry (Menozzi, Piazza and Cavalli-Sforza, 1978). For example, Novembre et al. (2008) famously found in a sample of Europeans that the first two PCs capture geographic variation in ancestry closely mirroring a map of Europe. We illustrate that PCs can capture real genetic structure in Figure 1, which plots the first four PCs from UK Biobank genetic data, with individuals labeled by self-reported ethnicity. Since self-reported ethnicity is correlated with genetic ancestry, the PCs cluster individuals by self-reported ethnicity. At the same time, the overlap across clusters highlights that self-identified ethnicity is not equivalent to genetic ancestry. Notably, many of the outliers—those falling far from the main clusters—self-report as mixed ancestry, suggesting that excluding such individuals may help reduce bias from population stratification. Based on these observations in many data sets, PCs quickly became the default tool for (restricting the sample and then) controlling for population stratification (Price et al., 2006).

For some time, it was widely believed that controlling for the top PCs—typically 10 or more—in a genetically **homogeneous** sample was a very effective way to guard against population stratification, at least when combined with other, standard quality-control measures (Bycroft et al., 2018; Winkler et al., 2014). But there is increasing recognition that **this strategy has** a number of limitations. First, the first few PCs capture only broad ancestry (e.g., north-south and west-east variation within Europe), missing finer population structure. For example, in Germany, cultural and marital patterns among Lutherans and Catholics may differ, yet these two groups can look nearly identical on top PCs. Adding more PCs (e.g., 100) is sometimes used as a fix, but as discussed next, we are skeptical that this strategy is effective. Second, large samples are needed to estimate more than the first few PCs accurately (Patterson, Price and Reich, 2006; Bloemendal, 2019). When PCs are estimated with noise, they do little to control for

Figure 1. Scatterplot of Principal Components in UKB



Note: Depiction of Principal Components in UK Biobank, by self-reported ethnicity.

stratification. Third, genetic PCs computed in the standard way do not capture recent population structure (i.e., occurring within the last few generations; see Zaidi and Mathieson, 2020, and Abdellaoui et al., 2022*a*). Controlling for PCs thus does not control for population stratification at the level of extended families and therefore does not address confounding from parental genetic effects or recent changes in population stratification. Therefore, even using well-estimated PCs in a large sample may leave substantial bias. Recent studies have confirmed these concerns, finding substantial bias in analyses even after controlling for PCs (e.g., Sohail et al., 2019; Berg et al., 2019; Lee et al., 2018). The findings of these papers are one impetus for the growing interest in family-based genetic studies.

D Family-Based GWAS with Parental Genotypes

A second control strategy leverages the natural experiment created by the random segregation of alleles during meiosis. This source of variation can be isolated by controlling for the parental genotypes (or their sum or midpoint). We refer to studies that control for parental genotypes as *family-based GWASs*.

At each genetic variant j, we denote the parental genotypes by $x_{f,j}$ (father) and $x_{m,j}$ (mother). The offspring's genotype x_j has conditional expectation:

$$\mathbb{E}\left[x_j \mid x_{f,j}, x_{m,j}\right] = \left(x_{f,j} + x_{m,j}\right)/2 \equiv x_{p,j},$$

where $x_{p,j}$ is the midpoint of parental genotypes. Therefore the offspring genotype can be expressed as:

$$x_j = x_{p,j} + x_{r,j},$$

where $x_{r,j}$ is a random deviation from the parental midpoint sometimes known as

the *Mendelian (or meiotic) segregation component.*[10] If Regression (4) is run with $x_{p,j}$ included among the controls,[11] the residual variation in x_j will isolate the random component $x_{r,j}$, ensuring that the β_j^{GWAS} estimator will have a causal interpretation.

This insight sets up a useful parallel for researchers familiar with quasi-experimental methods. As in other quasi-experimental designs, the estimand ends up being a weighted average of causal effects. Specifically, controlling for parental genotypes in the regression framework is formally equivalent to a two-stage least squares regression of y on x_j that uses $x_{r,j}$ as an instrument for x_j (see Appendix I). Therefore, as Veller, Przeworski and Coop (2024) showed, the estimate will be a *local* average treatment effect (LATE; see Imbens and Angrist, 1996). For each variant, only individuals **who have at least one heterozygous parent** contribute identifying variation. Moreover, across variants, the weighting will generally differ, depending on **how many heterozygous parents (zero, one, or both) an individual has** at locus j.

This framing highlights a conceptual connection to the broader econometric debate over the value of LATE estimates, with proponents (e.g., Imbens, 2010) emphasizing the value of a clearly defined causal estimand, recoverable under weak assumptions, whereas critics (e.g. Heckman and Urzúa, 2010) argue the estimand lacks generalizable policy relevance due to its dependence on a latent, instrument-specific subpopulation. In our context, the subpopulation is composed entirely of individuals with at least one heterozygous parent (which may be different for each SNP). Unlike in some econometrics contexts, however, here for many purposes the LATE is arguably of greater interest than the (equal-weighted) average treatment effect: because the LATE weights more heavily families with more variation in the inheritance of the allele, it is the weighted average that is most relevant to explaining the actual phenotypic variation. Moreover, for common SNPs in samples with relatively homogeneous genetic ancestry, we believe that heterogeneous effects by parental heterozygosity are likely to be minimal because parental heterozygosity at a single SNP provides very little information about heterogeneity in the environment or rest of the individual's genome. We therefore expect the LATE will generally be a close approximation to the average treatment effect, but we are unaware of any evidence on this issue.[12]

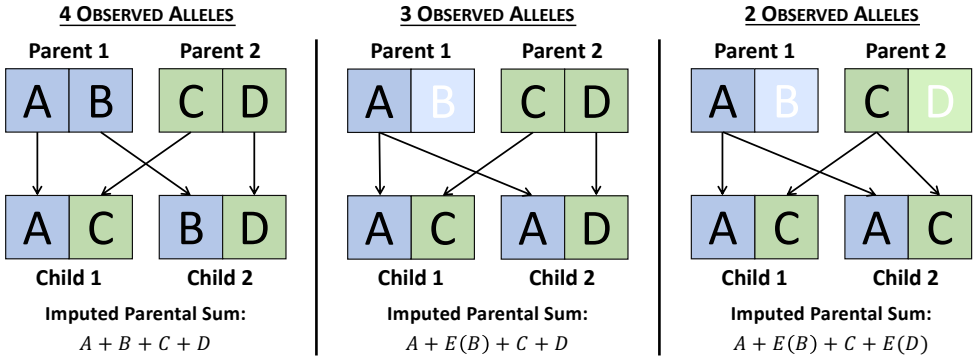
[10] A subtlety here is that $x_{r,j}$ is **not independent of, but** only mean independent of, $x_{p,j}$ and ϵ **(the residual from Equation (3))**. That is because at any genetic variant j, only heterozygous parents contribute to variation in the offspring genotype. Therefore, the variance of $x_{r,j}$ and the variance of the non-additive genetic factor, which is part of ϵ, both depend on the parental genotype $x_{p,j}$.

[11] Rather than controlling for the mean parental genotype, including the father's and mother's genotypes separately as controls in Regression (4) would also identify the self genetic effect, because $\text{span}\left(x_{p,j}\right) \subset \text{span}\left(x_{f,j}, x_{m,j}\right)$. In some cases, estimating the coefficients separately may be of substantive interest; however, neither the coefficients on the parental midpoint, nor the coefficients on mother's and fathers' genotypes when estimated separately, should be interpreted causally (Shen and Feldman, 2020; Young, 2023; Veller and Coop, 2024).

[12] The main case where we expect there may be a meaningful difference is when the sample has a mixture of genetic ancestries. In that case, parental heterozygosity can be informative about ancestry, which is correlated with environmental factors and other genotypes. However, most genetic analyses

Although we view the family-based GWAS identification strategy as the gold standard because it relies on a well-justified identifying assumption, it can still produce biased estimates if there are systematic sampling or attrition patterns that distort Mendelian segregation in the estimation sample. This issue is similar to the challenge of non-random attrition in randomized controlled trials, where inclusion in the analytic sample depends on both treatment assignment and potential outcomes. Importantly, however, even if there is selection in how parents are recruited, the estimates can still be internally valid as long as the *child's* genotype is conditionally independent of selection, conditional on the parents' genotypes. Of course, such selection still generates concerns about external validity.

Figure 2. Mendelian Imputation of Parental Genotypes (Young et al., 2022).



Note: A, B, C, and D refer to arbitrary alleles and are colored white if not transmitted to either child.

The main challenge for family-based GWAS so far has been the small number of samples available for analyses that have genotyped trios (both parents and their offspring). The scope for applying this identification strategy is rapidly broadening due to ongoing investments in new data collection, as well as improved methods that enable better use of existing data.

We mention here one of the most influential new methods because, beyond GWAS, it has become an important step in a number of social-science applications. Whenever the data set is missing trios but contains pairs of first-degree relatives who are genotyped, Young et al. (2022) shows how the Mendelian laws of inheritance can be used to impute missing parental genotypes.[13] Figure 2 illustrates the basic approach when applied to siblings, the most common scenario. The first imputation case is when all four alleles are observed in the offspring. In this situation, all four parental alleles can be recovered, and the parental genotype is known with

restrict the sample to have relatively homogeneous genetic ancestry (or separately analyze the sample by genetic ancestry), as mentioned in Section III.C.

[13]Hwang et al. (2020) proposed a related method, but it is less powerful because it does not use information about which SNPs are part of the same haplotype block. That information enables inferences about which alleles across SNPs were likely to have been inherited from the same parent.

certainty. In the remaining cases, only two or three of the parental alleles are observed in the children. In these situations, the expected mean parental genotype is imputed using sample allele frequencies in place of unobserved alleles. In a regression of the phenotype on child and imputed parent genotypes, the standard errors correctly account for the imputation uncertainty, and the coefficient on the offspring genotype is unbiased (provided that the imputation itself is unbiased; see Young et al., 2022).

E Sibling-Based GWAS with Family Fixed Effects

A third control strategy—*sibling-based GWAS*—is conceptually similar to a family-based GWAS and requires access to a sample with genotyped siblings.[14] Within each pair, we arbitrarily designate one individual as the index and assign the subscript *sib* to the other. For each SNP j, each sibling's outcome is regressed on their genotype at SNP j, controlling for family fixed effects. Thus, the approach can be viewed as a control strategy in which $\mathbf{z}$ is a vector of family fixed effects instead of parental genotypes. When all families consist of exactly two siblings, the fixed-effects estimator is equivalent to a sibling-difference estimator that regresses the difference in outcomes, $\Delta y = y - y_{sib}$, onto the difference in minor allele count, $\Delta x_j = x_j - x_{sib,j}$.

We highlight two subtle limitations of sibling-based approaches that, despite **their popularity**, are not widely appreciated (see Fletcher et al., 2024, for additional limitations). The first, highlighted by Young et al. (2022), concerns identification: the sibling-based approach only yields unbiased estimates of the self genetic effect, β_j^{GWAS}, if there are no sibling genetic effects (i.e., the genes of one sibling do not impact the outcome of the other sibling; see Section II.F). To see how the potential for bias arises, note that the identifying variation from the first sibling in a pair (the deviation of x_j from the family mean) can be expressed as:

$$x_j - \left(\frac{x_j + x_{sib,j}}{2}\right) = \frac{1}{2}\left(x_j - x_{sib,j}\right) = \frac{1}{2}\left(x_{r,j} - x_{r,sib,j}\right),$$

where $x_{sib,j}$ is the genotype of SNP j of the index individual's sibling and $x_{r,sib,j}$ is the random component of that genotype. If an individual's genotype affects their sibling, the sibling's genotype is contained within the residual of Equation (4). Thus, the expectation of the within-sibling-pair estimator is $\beta_j^{GWAS} - \beta_{sib,j}^{GWAS}$, where $\beta_{sib,j}^{GWAS}$ denotes the sibling genetic effect. Intuitively, the identifying variation comes from the deviation from the sibling mean, instead of the deviation from the parental midpoint, $x_{r,j}$. Thus, the estimated coefficient picks up both the effect of the individual having a higher genotype value and the effect of the sibling having a lower genotype value.

[14] Trejo and Kanopka (2024) develop a related, alternative approach that can be applied when only one sibling is genotyped, but it requires an assumption about the magnitude of genetic assortative mating.

The second limitation concerns power: even in the absence of sibling effects, the two-step approach that imputes parental genotypes and then controls for (imputed) parental phenotypes and family random effects (to control for shared family environment) yields more precise estimates than the sibling-difference estimator (Young et al., 2022).[15]

F Frontiers of GWAS

Nearly all GWASs performed to date have been population-based. Early GWASs, conducted in small samples of up to a few thousand individuals, were mostly underpowered and typically found little (e.g. Muglia et al., 2008; Wellcome Trust Case Control Consortium, 2007; Frayling et al., 2007; Stefansson et al., 2009). As sample sizes increased—up to roughly 5 million individuals in the most recent GWAS of height (Yengo et al., 2022) and roughly 3 million in the most recent GWAS of educational attainment (Okbay et al., 2022)—discoveries of genetic associations mounted, and these well-powered studies produced replicable results. In medical genetics, some GWAS results have led to biological insights and the identification of drug targets (Visscher et al., 2017). For social-science genomics, GWAS results have in some cases confirmed conclusions from twin studies and in other cases provided evidence about parameters that are typically not identified in twin studies, as we discuss in Section III.G below. For both medical and social-science applications, polygenic indexes are among the most important outputs of GWAS; we defer discussion of polygenic indexes to Section IV.

The reason why population-based GWASs have predominated—despite sibling and family-based GWAS having better identification—is that datasets with genotyped family members have been too small for adequate power.[16] Nonetheless, as more datasets with genotyped close relatives become available and as the limitations of population-based approaches are increasingly recognized, interest in

[15] The fundamental reason is that the sibling fixed-effects analysis, even in the absence of sibling effects, does not always exploit all the identifying variation. The two approaches are identical when siblings inherit either perfectly overlapping or non-overlapping alleles from their parents (the outer cases in Figure 2). However, when siblings inherit a common allele from one parent but distinct alleles from the other parent (the middle case in Figure 2), the common allele is double counted in the sibling mean, shading **the sibling mean** toward the observed sibling genotypes and reducing the variance of the identifying variation. Formal details are in Young et al. (2022).

[16] Not only are there fewer genotyped family-based samples but holding fixed the number of genotyped individuals, sibling and family-based GWASs are usually much less powerful than population-based GWASs. There are three main reasons for this. First, the bias in population-based GWAS may inflate the magitude of the estimates for some SNPs relative to family-based estimates. Second, the family- and sibling-based GWASs, by only leveraging within-family sources of genetic variation, have larger standard errors than a population-based GWAS. The within-family genetic variance is one half of the total genetic variance under random mating, and slightly less than half under realistic positive assortative mating. Third, if the study includes sibling pairs, some information is used to estimate the fixed or random effect for each sibling pair, reducing the degrees of freedom. (If siblings have correlated residuals, or if the parental genotypes explain a high proportion of the residual variance, then standard errors will be reduced. In practice, however, these effects only generate modest efficiency gains and are never large enough to offset the loss in power from the three factors described above.) Roughly speaking, a GWAS using parental genotypes as controls requires roughly twice the number of individuals (not including parents) than a population-based GWAS to obtain comparably sized standard errors, and a sibling-based GWAS **requires** even more.

family- and sibling-based approaches is growing. Two major efforts in this area are a sibling-based GWAS by Howe et al. (2022*b*) and a family-based GWAS by Tan et al. (2024), both analyzing a large number of phenotypes. By meta-analyzing results from multiple cohorts, these papers attain unprecedented sample sizes for their GWAS types (albeit very small relative to current population-based GWASs); for example, Howe et al. have a combined sample of up to 180,000 siblings for some phenotypes.

The results shed light on which results from population-based GWAS are most likely to be robust. In Howe et al. (2022*b*), for molecular phenotypes, such as low-density-lipoprotein cholesterol, the results were largely in line with previously reported findings in population-based GWASs. By contrast, for many of the social and behavioral phenotypes—educational attainment, age at first birth, number of children, cognitive ability, depressive symptoms, and smoking—the sibling-GWAS estimates of β_j^{GWAS} were smaller in magnitude on average than the population-based GWAS estimates. Tan et al. (2024) estimated the **coefficient-based** genetic correlation between educational attainment and several phenotypes using both population- and family-based summary statistics. A number of phenotypes—including height, lung health, and BMI—had a large estimate of genetic correlation with educational attainment using population-based GWAS results, but the family-based results were smaller in magnitude and statistically indistinguishable from zero. A plausible interpretation is that the controls used in the population-based GWASs failed to eliminate all confounding from factors that have correlated effects on educational attainment and these other phenotypes. We anticipate that the coming years will see many more sibling and family-based GWASs, and we will learn which conclusions drawn from population-based GWASs are robust and which will need to be updated.

Another ongoing development is diversification of study populations. As noted in Section III.C, population-based GWASs have traditionally been restricted to samples of relatively homogeneous genetic ancestry. The largest such samples have been from countries in Europe, the UK, the US, Australia, and New Zealand (Mills and Rahal, 2019), partly because these countries are wealthy and had the resources to fund large-scale genotyping efforts. As of June 2023, based on data from the GWAS Diversity Monitor (Mills and Rahal, 2020), an online database of GWASs, the average percentage of European-genetic-ancestry subjects in published studies is ∼95%, compared to their ∼15% share of the global population.

This "eurocentric bias" is widely considered to be a major problem (e.g., Martin et al., 2019; Duncan et al., 2019). Among various concerns, the most relevant for social-science applications is that, as we discuss in Section IV.C, the polygenic indexes constructed from existing GWAS results are less predictive among individuals with non-European genetic ancestries. This "limited portability" of polygenic indexes reduces their value.

Many efforts to mitigate Eurocentric bias are currently underway. Some of these are national biobanking efforts in some non-European-genetic-ancestry **coun-**

tries, with the largest samples to date being the China Kadoorie Biobank (a study with ∼500,000 genotyped individuals currently), Biobank Japan (∼200,000 genotyped individuals currently), and the Taiwan Biobank (∼140,000 genotyped individuals currently). In the U.S., initiatives such as the Million Veterans Project, All of Us, the Multi-Ethnic Study, the UCLA Atlas Biobank, and the Together for CHANGE initiative are collecting relatively large minority samples. The Pan UKB Project has analyzed data from the UK Biobank for individuals with non-European genetic ancestries that would normally be discarded. Direct-to-consumer genetic testing companies, despite having a disproportionately European-genetic-ancestry customer base, nonetheless have many non-European-genetic-ancestry customers who have consented to participate in research. Some of these companies, such as 23andMe, are helping to mitigate Eurocentric bias by contributing to GWASs in diverse samples (e.g., Yengo et al., 2022). Funding agencies, including the U.S. National Institutes of Health, have prioritized collecting and analyzing genetic data from non-European-genetic-ancestry samples. Major journals for genetics research have prioritized publishing such work.

Unfortunately, the genetic-data-collection efforts in developing countries remain small. This hampers genetics research since populations in these countries, especially in Africa, harbor a large share of the global genetic diversity (Mills and Rahal, 2019).

G Estimating Heritability and Genetic Correlation

In most social-science applications, researchers are less interested in the contribution of a single SNP to an outcome or model and more concerned with the role of the genetic factor as a whole. Accordingly, several methods have been developed to estimate heritability and genetic correlation (see Section II.D for formal definitions and discussion). These estimators rely on different assumptions and target different estimands (e.g., narrow- vs. broad-sense heritability). We provide a high-level overview of the main approaches below.

Historically, in the absence of molecular genetic data, estimates of heritability and genetic correlation relied on family-based designs, such as classical twin, family, and adoption studies. This line of work, introduced into economics by Taubman (1976), has been previously reviewed for economic audiences (e.g., Beauchamp et al., 2011; Benjamin et al., 2012; Sacerdote, 2011). These methods exploit *expected* degrees of genetic and environmental similarity to identify structural parameters. For example, assuming that monozygotic twins raised apart have identical genetic factors but uncorrelated residuals, the correlation of their phenotypes can be interpreted as an estimate of broad-sense heritability. As additional types of relative pairs are introduced into the analysis, more complex models with more parameters can be identified. However, these approaches depend on strong assumptions, and different sets of plausible assumptions (e.g., regarding assortative mating, gene–environment correlation, or the contribution of non-additive

genetic components) can yield **conflicting** estimates (see Loehlin, 1978).

Modern genomic data have enabled approaches that rely on weaker assumptions (including extensions of earlier designs; see, e.g., Beauchamp et al., 2023). There are two classes of such methods: genomic-relatedness methods and GWAS-based methods. Genomic-relatedness methods typically require large samples of individual-level, linked genotype and phenotype data. These methods infer heritability or genetic correlation by comparing phenotypic similarity between individuals with their *measured* genetic similarity. Due to the same quasi-experimental logic and with the same caveats as in Section III.B, when researchers control for parental genotypes (Young et al., 2018) or focus solely on comparisons among siblings (Visscher et al., 2006; Haseman and Elston, 1972; Markel et al., 2025), these estimators **have credible causal identification**. However, just as most GWASs have been population-based, most genomic-relatedness methods have been applied in non-family samples, following Yang et al. (2010). In that case, for identification these estimators rely on controlling for genetic principal components and are therefore sensitive to the confounds discussed in Section III.B.

GWAS-based methods use only GWAS summary statistics and a modest amount of genomic data from an ancestry-matched "reference panel" that is used to provide an estimate of the population's LD matrix. These methods have become more common due to their reduced data requirements (no individual-level data) and relative computational ease. Several such approaches exist (e.g., Speed and Balding, 2019), but the most widely used is LD Score regression (Bulik-Sullivan et al., 2015*b*,*a*). Using the reference panel, this method computes an LD score for each SNP, which is a measure of how strongly that SNP is correlated with other SNPs across the genome. For heritable phenotypes, SNPs with higher LD scores tend to have larger GWAS associations because they are expected to **be correlated with** more causal variants. This relationship is used to infer heritability and genetic correlation. However, because these estimates depend on GWAS summary statistics, they inherit the biases of those statistics—for instance, biases arising from gene–environment correlation in population-based GWAS. These biases can be eliminated if family-based GWAS summary statistics are used, which we expect will become more common once family-based GWAS become sufficiently powered.

Genomic-data-based estimators have helped resolve longstanding debates that could not be settled using classical twin, family, and adoption studies. For example, in the case of educational attainment, there is now strong evidence from genomic data for both gene–environment correlation (Young et al., 2018) and assortative mating (Robinson et al., 2017; Lee et al., 2018). While some classical models permitted a substantial role for non-additive genetic components, genomic data have ruled out dominance variance (a type of non-additive genetic variance) as an important source of variation in educational attainment (Okbay et al., 2022). These findings affirm Goldberger's caution, directed at the classical twin and family studies, that misspecified models may yield biased estimates despite good statistical fit—and that relying on goodness-of-fit tests alone to assess

model validity is misguided (Goldberger, 1978, p. 72).

IV Polygenic Indexes

Most applications using genetic data in the social sciences use polygenic indexes (PGIs). Although PGIs had been discussed earlier (Wray, Goddard and Visscher, 2007), the first paper in humans genetics to construct and analyze a PGI was a GWAS of schizophrenia published in 2009 (Purcell et al., 2009). Since then, PGIs have been increasingly used in research related to the genetics of behavioral phenotypes (Becker et al., 2021).

There are two main reasons for the use of PGIs in social-science research, one statistical and one conceptual. Statistically, because PGIs generally explain much more variance than individual SNPs, analyses using a PGI will generally have much greater statistical power. Conceptually, the PGI is an empirical proxy for the additive SNP factor and thus captures the combined explanatory power of measured SNPs. In this section, we discuss PGIs, their predictive power, and their appropriate interpretation.

A PGI Definition and Interpretation

In general, we define a PGI as a standardized, weighted sum of the genotypes of a set of measured genetic variants:

$$g_{\mathbf{w}} \equiv \frac{\tilde{\mathbf{x}}\mathbf{w}}{\text{std}\left(\tilde{\mathbf{x}}\mathbf{w}\right)},$$

where $\tilde{\mathbf{x}}$ is the vector of measured genotypes, $\mathbf{w}$ is a vector of weights (the "PGI weights"), and $\text{std}\left(\cdot\right)$ takes the standard deviation of its argument across a population of individuals. Typically, the measured genotypes are SNPs, and the PGI weights are chosen with the goal of having the PGI approximate the standardized additive SNP factor for some phenotype (**where the additive SNP factor is as** defined in Section II.C) as well as possible.[17]

The PGI would equal the standardized additive SNP factor if $\mathbf{w} = \tilde{\boldsymbol{\beta}}$, where $\tilde{\boldsymbol{\beta}}$ is the vector of population regression coefficients defined in Equation (2). In practice, researchers cannot set $\mathbf{w}$ equal to $\tilde{\boldsymbol{\beta}}$ because the SNP coefficients are estimated, not known, as we discuss below. In addition, in most applications to date researchers have used SNP coefficients estimated from a population-based GWAS with imperfect controls $\mathbf{z}$, and consequently, these coefficients may be biased. To allow **us to analyze PGIs constructed with such biased coefficients**, we define the *optimal predictor weights* $\breve{\boldsymbol{\beta}}$ given some set of controls $\mathbf{z}$ by

[17]The discussion and analysis in this subsection apply also to PGI weights chosen such that the PGI approximates some other quantity. For example, the genetic principal components estimated from a sample (discussed in Section III.C) are PGIs that aim to approximate the population's (true) genetic principal components.

the following population **least-squares projection**:

$$\left\{\breve{\boldsymbol{\beta}}, \breve{\boldsymbol{\gamma}}\right\} \equiv \underset{\mathbf{b},\mathbf{a}}{\operatorname{argmin}} \left\{\mathbb{E}\left[\left(y - \tilde{\mathbf{X}}\mathbf{b} - \mathbf{Z}\mathbf{a}\right)^2\right]\right\} \tag{6}$$

for a certain phenotype y, **where $\tilde{\mathbf{X}}$ and $\mathbf{Z}$ are the random variables from which the observed genotype vectors and control-variable values are drawn and the expectation is taken over their joint distribution.** We refer to $\breve{y} \equiv \tilde{\mathbf{x}}\breve{\boldsymbol{\beta}}$ as the *optimal predictor*. If the controls $\mathbf{z}$ are sufficient for $\tilde{\mathbf{x}}$ to be as good as random conditional on $\mathbf{z}$, then $\breve{\boldsymbol{\beta}}$ is equal to the additive SNP factor weights $\tilde{\boldsymbol{\beta}}$, and the optimal predictor is equal to the additive SNP factor; otherwise, they **are generally not** equal. Parallel to the notion of SNP heritability from the causal model, we define $\breve{h}^2 \equiv \operatorname{Var}(\breve{y})$ as the *optimal predictive power* for y since it is the maximal predictive power for y from a linear combination of observed SNPs given the set of controls. One way to estimate the optimal predictive power is using LD score regression on GWAS summary statistics that use $\mathbf{z}$ as the set of controls.

The sample analog of Equation (6) cannot be estimated for the same multicollinearity and short-rank reasons discussed in Section III. In this context where the goal is prediction (rather than estimation of SNP effects), most of the commonly used estimators for $\breve{\boldsymbol{\beta}}$ are Bayesian. They take as inputs a set of GWAS summary statistics, an estimate of the LD matrix $\boldsymbol{\Sigma}$ obtained from some reference sample, and a Bayesian prior distribution of effect sizes.[18] The estimators adjust the GWAS estimates to take into account correlation across SNPs, as captured by the LD matrix, and shrink them toward the prior. Specifically, the estimators set each SNP's PGI weight equal to the mean of its Bayesian posterior-effect distribution; the estimators differ from each other mainly in their assumptions about the prior distribution and, for computational tractability, in the assumptions and approximations they make about the LD matrix, $\boldsymbol{\Sigma}$ (e.g., Vilhjálmsson et al., 2015; Ge et al., 2019; Zhang et al., 2021; Lloyd-Jones et al., 2019). These differences affect finite-sample performance and computational speed but do not matter for the purposes of discussion here.

We denote the resulting weights by $\hat{\boldsymbol{\beta}}$ and the corresponding PGI by

$$\hat{g} \equiv \frac{\tilde{\mathbf{x}}\hat{\boldsymbol{\beta}}}{\operatorname{std}\left(\tilde{\mathbf{x}}\hat{\boldsymbol{\beta}}\right)}.$$

[18] A simpler approach, developed earlier and still widely used (especially in medical applications), is called "pruning and thresholding." In this approach, the PGI is constructed from a set of approximately mutually uncorrelated ("pruned") SNPs whose GWAS p-value is below some threshold, and their PGI weights are set equal to their GWAS estimates. For highly polygenic phenotypes—including social and behavioral phenotypes—pruning-and-thresholding makes less sense than approaches that use all the measured SNPs because all the measured SNPs could add information to the PGI. In addition to Bayesian approaches and pruning-and-thresholding, machine-learning approaches also exist (e.g., Widen et al., 2021; Zhao et al., 2021).

Following Tucker-Drob (2017)) and Becker et al. (2021), the remainder of this section explains the interpretation of the PGI $\hat{g}$: it is a (standardized) noisy measure of the optimal predictor $\breve{g}$, with classical measurement error. This result is central to the conceptual appeal of the PGI, and **because the measurement error is classical,** there are relatively straightforward methods of correcting for **the** measurement error in applications (as discussed in Section IV.E).

To keep focus on the central issues, we assume that the LD-estimation sample grows at the same rate as the GWAS sample (as would be true, for example, if the LD matrix were estimated in the GWAS sample itself). We also assume that the GWAS sample, the LD-estimation sample, and the prediction sample are all drawn from the same population, except in Section IV.C below, where we discuss the implications of using PGI weights based on a population that is different from the population of the prediction sample.

As noisy estimates of the optimal predictor weights, the PGI weights can be expressed as $\hat{\boldsymbol{\beta}} = \breve{\boldsymbol{\beta}} + \mathbf{u}$ for some **sampling-error** vector $\mathbf{u}$. The PGI $\hat{g}$ is thus a (standardized) noisy measure of the optimal predictor $\breve{g} \equiv \tilde{\mathbf{x}}\breve{\boldsymbol{\beta}}$:

$$\hat{g} = \frac{\tilde{\mathbf{x}}\hat{\beta}}{\text{std}\left(\tilde{\mathbf{x}}\hat{\boldsymbol{\beta}}\right)} = \frac{\tilde{\mathbf{x}}\breve{\boldsymbol{\beta}} + \tilde{\mathbf{x}}\mathbf{u}}{\text{std}\left(\tilde{\mathbf{x}}\breve{\boldsymbol{\beta}} + \tilde{\mathbf{x}}\mathbf{u}\right)} = \frac{\breve{g} + e}{\text{std}\left(\breve{g} + e\right)},$$

where $e \equiv \tilde{\mathbf{x}}\mathbf{u}$ is noise that comes from the sampling error $\mathbf{u}$. If $\hat{\boldsymbol{\beta}}$ were estimated by the sample analog of the population regression Equation (6) above (which is not feasible), then the noise e would be mean zero, uncorrelated with the optimal predictor $\breve{g}$, and independent of all variables in any independent prediction sample. Moreover, it follows from $\text{Cov}\left(\breve{g}, e\right) = 0$ that $\text{Var}\left(\breve{g} + e\right) = \text{Var}\left(\breve{g}\right) + \text{Var}\left(e\right)$. For a standard Bayesian approach to constructing a PGI discussed above, these properties do not hold, but they hold approximately if the GWAS sample size (the sample size underlying $\hat{\boldsymbol{\beta}}^{\text{GWAS}}$ and the estimated LD matrix) is large. Becker et al. (2021, see Supplementary Materials 4) derives formulas for these approximations and calculates that the approximations are tight for the PGI derived from a recent GWAS of educational attainment (Lee et al., 2018). When these approximations are tight, e can be treated as classical measurement error.

This result that the measurement error is classical may be surprising. One might have had the intuition that the measurement error would be non-classical because the PGI coefficients $\hat{\boldsymbol{\beta}}$ are estimated less precisely for some SNPs (rarer SNPs, which have less genotypic variation) than others. If the SNPs' *genotypes* were measured with different amounts of error, the measurement error would indeed be non-classical, but different amounts of measurement error in the PGI weights do not cause the measurement error in the PGI to be non-classical.

B PGI Predictive Power

In this subsection, we derive an analytic formula for the predictive power of a PGI. In some applications, including clinical use of PGIs to assess disease risk (e.g., Khera et al., 2018), the predictive power of a PGI is central to its usefulness. In social-science research applications, the predictive power of a PGI is a central factor in the statistical power of the analysis. Statistical power calculations are valuable both for deciding which analyses to undertake and for evaluating the credibility of findings (Bayarri et al., 2016; Maniadis, Tufano and List, 2014). For these reasons, we believe that the formula can be useful to social scientists.

We focus on a univariate regression of a phenotype y_{pred} on the PGI $\hat{g}$ constructed to predict a possibly different phenotype y_{GWAS}, and we briefly discuss afterward how covariates complicate the analysis. Our measure of predictive power is the coefficient of determination (R^2) from a population regression. We assume that the prediction population is independent of the GWAS population used to estimate the PGI weights (for phenotype y_{GWAS}). For now, we assume the two populations have a common LD matrix (e.g., they are randomly sampled from the same population). We denote the optimal predictive power for y_{GWAS} in the GWAS population by $\breve{h}^2_{GWAS}$ and the optimal predictive power for y_{pred} in the prediction population by $\breve{h}^2_{pred}$. Following the derivation in Daetwyler, Villanueva and Woolliams (2008) and generalizations in de Vlaming et al. (2017) and Okbay et al. (2022), we show in Appendix II:

$$R^2 = \left(\breve{h}^2_{pred} r^2_{\mathbf{x}\boldsymbol{\beta}}\right)\left(\frac{\breve{h}^2_{GWAS}}{\breve{h}^2_{GWAS} + M/N}\right), \tag{7}$$

where $r_{\mathbf{x}\boldsymbol{\beta}}$ is the correlation between the optimal predictor for y_{GWAS} in the GWAS population and the optimal predictor for y_{pred} in the prediction population, a type of "genetic correlation" as defined in Section II.D; M is a constant; and N is the GWAS sample size underlying the PGI weights.

The first term in Equation (7)—the optimal predictive power for y_{pred} in the prediction population, $\breve{h}^2_{pred}$, multiplied by $r^2_{\mathbf{x}\boldsymbol{\beta}}$—is the predictive power that would be achieved if the PGI weights were estimated from an infinite GWAS sample ($N \to \infty$). We call it the *asymptotic-R^2 term*. The parameter $\breve{h}^2_{pred}$ could be larger or smaller than $\breve{h}^2_{GWAS}$, depending on, among other things, what y_{pred} and y_{GWAS} are (e.g., cognitive performance may be more predictable than educational attainment), how they are measured (e.g., the amount **of** measurement error), and what the GWAS and prediction populations are. The attenuation factor $r^2_{\mathbf{x}\boldsymbol{\beta}}$ is bounded above by one—a bound achieved when, for example, the phenotype and populations in the prediction and GWAS samples are identical—and will be smaller than one to the extent that the optimal predictors for y_{pred} and y_{GWAS} differ.

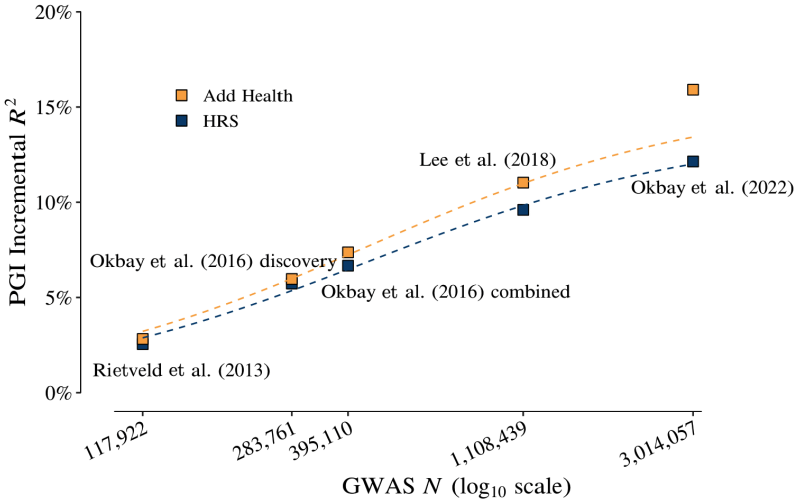
The second term in Equation (7)—which we call the *estimation-precision term*—is between 0 and 1 and is related to the signal-noise ratio in the GWAS: the optimal predictive power $\breve{h}^2_{GWAS}$ is a measure of the signal, and M/N is a measure of the noise. The constant M depends on the LD matrix. Under our assumption that the LD matrix is full rank, M is equal to the number of SNPs in the PGI. Otherwise, M is smaller than that number. Using population-genetic theory, some simplifying assumptions, and estimates of related quantities, M has been roughly estimated to be 60,000 to 70,000 in European-genetic-ancestry populations (Hayes, Visscher and Goddard, 2009; Rietveld et al., 2013; Wray et al., 2013). Alternatively, M can be estimated by fitting Equation (7) based on the N's of previous GWAS, the R^2's of the resulting PGIs, and an estimate of $r^2_{\mathbf{x}\boldsymbol{\beta}}$ (as in Okbay et al., 2022). After either calibrating $M \approx 70,000$ or estimating M from previous GWASs, Equation (7) can be used to forecast the predictive power of a PGI from a future GWAS with a larger sample size.

In practice, when researchers examine the predictive power of a PGI, they most commonly report the *incremental* R^2: the change in R^2 from adding the PGI to a regression of the phenotype on a baseline set of covariates. These baseline covariates are typically the same as those included in a GWAS: age, year of birth, and genetic principal components. To illustrate, Figure 3 shows how, for each of two prediction datasets, the incremental R^2 of the PGI for educational attainment has increased as GWAS discovery samples have increased from roughly 100,000 individuals to roughly 3,000,000 individuals. The two prediction datasets are the Health and Retirement Study (HRS), a U.S. nationally representative sample of older Americans, and the National Longitudinal Adolescent to Adult Health Study (Add Health), a U.S. nationally representative sample of younger Americans. In both datasets, the predictive power of the PGI is increasing in the GWAS sample size. At each sample size, the predictive power appears to be larger in Add Health.

Equation (7) is derived for a univariate regression, rather than for the incremental R^2 between two multivariate regressions, but it can nonetheless provide some useful insights. For example, the equation implies that the difference in predictive power across the datasets is due to a difference in $\breve{h}^2_{pred} r^2_{\mathbf{x}\boldsymbol{\beta}}$; Okbay et al. (2022) indeed estimate a larger optimal predictive power $\breve{h}^2_{pred}$ in Add Health, albeit with large standard errors. The dashed line **in Figure 3** fits Equation (7) to the points in the figure separately for each dataset. The functional form implied by Equation (7) provides a good fit, except for the predictive power in Add Health from the most recent GWAS, which is larger than expected.

As another empirical illustration of Equation (7), Mostafavi et al. (2020) study PGIs for diastolic blood pressure, BMI, and educational attainment and document how their predictive power varies with the sex, age, and socioeconomic status of the prediction sample. Consistent with Equation (7), for each PGI separately, Mostafavi et al. (2020) find that the incremental R^2 is larger when the GWAS

Figure 3. Predictive Power of PGI for Educational Attainment as a Function of Sample Size



Note: The x-axis is the sample size of the GWAS on a log scale. The y-axis is the incremental R^2 of the EA PGI constructed from the GWAS summary statistics, in each of two prediction samples independent of the samples used in the original GWAS meta-analysis. Incremental R^2 is the increase in R^2 after adding the PGI to a regression of years of schooling on the following controls: a full set of dummy variables for year of birth, an indicator variable for sex, a full set of interactions between sex and year of birth and the first ten genetic principal components. Figure is adapted from Okbay et al. (2022).

sample is demographically more similar to the prediction sample (implying higher $r^2_{\mathbf{x}\boldsymbol{\beta}}$) and when the optimal predictive power is larger in the prediction sample.

Although an incremental R^2 relative to a baseline set of covariates can be a useful measure, it may be a misleading measure of the gain from including the PGI in social-science research for two reasons. First, social-science applications usually include a richer set of covariates that absorb more of the variation explained by the PGI. To illustrate this point, Lee et al. (2018, see their Supplementary Figure 12(b)) report how the incremental R^2 of the PGI for educational attainment declines as additional covariates are included in the regressions. With just the baseline covariates of age, sex, and genetic principal components, the incremental R^2 is roughly 11%, but it falls to roughly 5% when additionally controlling for marital status, income, mother's education, and father's education. Second, the incremental R^2 may not be the relevant measure of predictive power, depending on the purpose of the analysis. For example, following Rietveld et al. (2013), **we show in Appendix IV** that, for the purpose of increasing the precision of a treatment effect estimate, the efficiency gains from adding a control for a PGI can be substantial, even when—indeed, *especially* when—the remaining set of covariates explain much of the variation. In that context, the relevant measure of predictive power is the R^2 from a regression of the residual (of the outcome after controlling for the covariates) on the PGI. This R^2 is larger when the covariates explain more variation.

C PGI Portability Across Populations

Many potential social-science applications involve populations such as African Americans, Hispanic, and other groups whose genetic ancestry is not the same as that of contemporary European populations. One major limitation of PGIs is that, at present, their predictive power is typically much lower in such populations. In the literature, this issue is called the problem of limited "portability."

Theoretically, relative to the issues already discussed, the new issue introduced when the GWAS and prediction samples consist of people with different genetic ancestries is that the LD matrix in the GWAS sample is not equal to the LD matrix in the prediction sample. In Appendix II, we generalize Equation (7) to derive an exact formula for the predictive power **in this case** that is novel as far as we are aware. Related approximate formulas are derived in Wientjes et al. (2015; 2016), Wang et al. (2020) and Ding et al. (2023). Our formula shows that the difference in LD matrices generates three effects. First, the squared genetic correlation parameter $r^2_{\mathbf{x}\boldsymbol{\beta}}$ is reduced because the optimal-predictor weights from the GWAS sample are no longer optimal when applied in the prediction sample. Second, the estimation-precision term is smaller because the SNPs that have the largest genotypic variance in the GWAS sample—which are the SNPs that contribute most to prediction accuracy in the GWAS sample—are those whose optimal predictor weights are estimated most precisely, but those SNPs may not be the ones that have the largest genotypic variance in the prediction sample. Both

of these effects unambiguously reduce R^2. Third, the estimation-precision term is modified, due to differences between the samples in the frequencies of alleles that have larger coefficients in predicting the phenotype. This third effect could go in either direction, but under reasonable assumptions for most applications, the effect is small in expectation.

In practice, the GWASs underlying PGI weights are typically conducted in samples of individuals of European genetic ancestries (see Section III.F)—and empirically, on average across phenotypes, and for almost all phenotypes that have been studied, PGIs have less predictive power in samples of non-European-genetic-ancestry individuals. For example, Martin et al. (2019) find that, on average across 17 anthropometric and blood phenotypes, relative to the PGI R^2 in European-genetic-ancestry samples, the R^2 is roughly 33% smaller in native American and South Asian genetic-ancestry samples, roughly 50% smaller in East Asian genetic-ancestry samples, and roughly 75% smaller in African genetic-ancestry samples (similar results are reported in Duncan et al., 2019, Martin et al., 2017, and Alemu et al., 2025*a*). The decline in the PGI R^2 from European-genetic-ancestry to African-genetic-ancestry populations is even more pronounced for educational attainment, roughly 85% (Lee et al., 2018; Okbay et al., 2022).

Consistent with theoretical expectations, the average decline in predictive power tracks qualitatively with genetic distance from European genetic ancestry (and indeed, even among individuals with European genetic ancestry, average predictive power of a PGI is lower for individuals more distantly related to the GWAS sample; Ding et al., 2023). To estimate quantitatively the extent to which differences in LD matrices explain the decline in predictive power for various phenotypes, Wang et al. (2020) used an approximate R^2 formula, together with estimates of LD matrices from different populations and GWAS results for eight anthropometric and health-relevant phenotypes. They find that 70%-80% of the drop in PGI R^2 from European-genetic-ancestry to African-genetic-ancestry populations can be accounted for by the LD-matrix differences.

In the long term, constructing more predictive PGIs in non-European-genetic-ancestry populations will become possible as more genotyped samples from those populations become available. In those larger samples, GWASs can be conducted, and population-specific PGI weights can be obtained. In the shorter term, new statistical methods can partially substitute for larger GWAS samples (Turley et al., 2021; Ruan et al., 2022; Miao et al., 2022*b*). These methods leverage results from large-scale GWASs in European-genetic-ancestry populations to create synthetic GWAS results for other populations, using the populations' LD matrices to "translate" GWAS associations across populations.

The discussion above has focused on the problem of using PGIs trained in one population to predict phenotypic variation *within* a population with different genetic ancestry. Additional challenges arise when comparing the *level* of a PGI across individuals from different populations. Even when the two populations are genetically similar, such comparisons can be confounded by different mean levels

of the phenotype (for non-genetic reasons), different true genetic effects across the populations (most notably due to gene-environment interactions), different patterns of gene-environment correlation, and different prediction-error variances. When the populations are from different genetic ancestries, the non-genetic differences may be greater, and since the LD matrices differ, the SNPs included in the PGI will capture causal effects (including those of unmeasured genetic variants) to different degrees, exacerbating these challenges. Indeed, comparisons of PGI levels across populations with different genetic ancestry are unlikely to be valid in most cases (unless the ancestries are sufficiently similar). Martin et al. (2017) (their Figure 4A) provided a striking empirical example: they compared the distributions of height PGIs for several different populations with different genetic ancestries using data from the 1000 Genomes Project. They found that the African populations sampled are genetically predicted to be considerably shorter than all the European populations sampled, which contradicts empirical observations on measured height.

D Estimating and Interpreting the "Causal Effect of a PGI"

In most applications, researchers are interested in **studying** causal effects of genetic influences. The **leading** empirical tool available to study such genetic influences is a PGI, and one might naturally anticipate that regressing an outcome on a PGI, controlling for parental PGIs, will identify causal genetic effects. In this subsection, we will interpret the estimand from such a regression. Our results will show that while the estimand does indeed capture causal genetic effects, there are some important subtleties that researchers should be aware of.

The first issue is that the notion of the "causal effect of a PGI" is not conceptually straightward. We put "causal effect of a PGI" in quotes because we do not have in mind a hypothetical experiment in which we examine the effect of changing the PGI. As Veller, Przeworski and Coop (2024) explain, formulating such a hypothetical experiment is challenging, for two reasons.[19] First, the PGI weights typically do not represent causal effects of the SNPs included in the PGI. Even if the PGI weights were obtained from a family-based GWAS, the PGI weight on a SNP partly reflects the causal effects of unmeasured genetic variants that are correlated with measured SNPs. (For this same reason, in Section II.C, we could not express the additive SNP factor in terms of potential outcomes.) Second, the PGI is an index. Thus, even if the PGI weights were the causal effects of the included SNPs, which SNPs' genotypes were changed when considering a hypothetical experiment of changing the PGI by some amount could matter. Thus,

[19] Perhaps the most compelling hypothetical experiment would be to imagine that prospective parents create many embryos, one of which is chosen at random and results in a live birth. Each embryo has a different genotype vector randomly assigned (conditional on the parents) and therefore a different PGI. The association between the PGIs and the potential outcomes has a causal interpretation—but it is an average treatment effect conditional on the parents, and it is not clear how it relates to a quantity that could be estimated. Moreover, it does not solve the two conceptual challenges to defining the "causal effect of a PGI" described in this paragraph.

two individuals whose PGIs changed by the same amount might have different hypothetical experiments that define the effect.[20]

Instead, like Veller, Przeworski and Coop (2024), we work backwards from the "causal effect of a PGI" that is estimable and provide an interpretation of the estimand. Veller, Przeworski and Coop (2024) derive the coefficient from a regression of sibling differences in the phenotype on sibling differences in the PGI and show that, under the assumption that there are no sibling genetic effects, this is equal to the coefficient on the individual's PGI from a regression of the individual's phenotype onto their PGI and that of their parents. We instead directly derive the coefficient from latter regression that includes parental PGIs, which allows us to additionally derive expressions for the coefficients on the parental PGIs.

To begin, extending the single-SNP estimation framework from Section III.D, an individual's genotype vector, $\mathbf{x}$, can be decomposed into the mean parental genotype vector, $\mathbf{x}_p \equiv \frac{\mathbf{x}_f+\mathbf{x}_m}{2}$, and a random deviation, $\mathbf{x}_r \equiv \mathbf{x} - \frac{\mathbf{x}_f+\mathbf{x}_m}{2}$. Consider the population regression of some outcome variable y on $\mathbf{x}$ and $\mathbf{x}_p$,

$$y = \mathbf{x}\boldsymbol{\beta} + \mathbf{x}_p\mathbf{b}_p + \xi. \tag{8}$$

As in the single-SNP case, the coefficient on the child's genotype vector, $\boldsymbol{\beta}$, is the vector of causal genetic effects, though the coefficient on the parent's genotype vector, $\mathbf{b}_p$, generally does not have a causal interpretation. Note that for what follows, whether y is the phenotype corresponding to the PGI or some other outcome makes no difference.

Using the same decomposition of the genotype vector, any PGI with weight vector $\mathbf{w}$ can be expressed as

$$g_{\mathbf{w}} \equiv \mathbf{x}\mathbf{w} = \mathbf{x}_r\mathbf{w} + \mathbf{x}_p\mathbf{w}$$

where (unlike in Section IV.A above) we now express the PGI as a weighted sum of the genotypes of *all* genetic variants, with $w_j = 0$ for every unmeasured genetic variant j. Also, to reduce notational clutter, we assume $\mathbf{w}$ has been rescaled to standardize the PGI: $\text{std}(\mathbf{x}\mathbf{w}) = 1$. Now consider a population regression of y on the individual's PGI $g_{\mathbf{w}}$ and the sum of parental PGIs, $p_{\mathbf{w}} \equiv \mathbf{x}_f\mathbf{w} + \mathbf{x}_m\mathbf{w}$:

$$y = \alpha_g g_{\mathbf{w}} + \alpha_p p_{\mathbf{w}} + u \tag{9}$$

(where we define $p_{\mathbf{w}}$ as a sum rather than an average because it makes the ex-

[20] Which genotypes were changed would not matter if two conditions are both satisfied: (i) the additive model is true (i.e., the additive genetic factor coincides with the genetic factor), and (ii) the analysis focuses on the phenotype corresponding to the PGI (for example, an analysis of educational attainment using the PGI for educational attainment). While (i) may often be a reasonable approximation, most social-science applications of PGIs violate (ii). For example, consider a study of the effect of the PGI for educational attainment on income (as in Papageorge and Thom, 2020). To see how it may matter which genotypes were changed, suppose changing either of two SNPs would increase the PGI by one unit. If one of the SNPs affects income and the other does not, the two hypothetical experiments have different effects.

pressions for α_p and α_g symmetric). In Appendix III, we derive the relationship between the coefficients from the regressions in equations (8) and (9). Although our analysis there is more general and the resulting formulas correspondingly more complex, here we present the results under the assumption of a randomly mating population:

$$\alpha_g = \frac{\mathbf{w}'\boldsymbol{\Sigma}\boldsymbol{\beta}}{\mathbf{w}'\boldsymbol{\Sigma}\mathbf{w}} \tag{10}$$

$$\alpha_p = \frac{\mathbf{w}'\boldsymbol{\Sigma}\mathbf{b}_p}{\mathbf{w}'\boldsymbol{\Sigma}\mathbf{w}}, \tag{11}$$

where $\boldsymbol{\Sigma} \equiv \mathrm{Var}(\mathbf{x}) = \mathrm{Var}(\mathbf{x}_f) = \mathrm{Var}(\mathbf{x}_m)$ is the LD matrix, which is the same in the parents and children due to the random-mating assumption.

Equation (10) is equal to (but formulated differently than) the expression derived by Veller, Przeworski and Coop (2024). They show that this coefficient does *not* correspond to any weighted sum of the individual-specific causal effects of genetic variants, for any fixed weights across individuals. They conclude that "the family-based estimate is a strangely weighted average across [genetic variants] and across families." Our formulation of Equation (10) makes clear that the coefficient on the PGI, α_g, is a weighted sum of *average* (across individuals) causal effects: the β_j's. For this reason, we contend that it is justified to refer to the coefficient on the PGI as having a causal interpretation.

What exactly is the estimand α_g? Our formulation shows that it is the coefficient from a regression of the additive genetic factor for the outcome y on the PGI, estimated using only the random component of the genotype vector. For example, if y is income and the PGI is for educational attainment, then α_g measures the slope of the relationship between a change in the PGI (caused by exogenous changes in genotype) and a change in the additive genetic factor for income. This is most straightforward to see if we ignore the parental PGI in regression Equation (9) and treat the genotype vector as exogenous. Then, regressing the additive genetic factor, $\mathbf{x}\boldsymbol{\beta}$, on the PGI gives coefficient

$$\frac{\mathrm{Cov}(\mathbf{x}\mathbf{w}, \mathbf{x}\boldsymbol{\beta})}{\mathrm{Var}(\mathbf{x}\mathbf{w})} = \frac{\mathbf{w}'\boldsymbol{\Sigma}\boldsymbol{\beta}}{\mathbf{w}'\boldsymbol{\Sigma}\mathbf{w}},$$

which is precisely α_g. Equivalently, α_g can be understood as the slope from a *generalized* least squares regression **of the β_j's on the PGI weights** w_j, where the dispersion matrix is the LD matrix. The LD matrix weights more heavily SNPs whose genotypes have higher variance (i.e., more common SNPs) and SNPs that **are more strongly correlated with** other genetic variants (included unmeasured variants) because these SNPs are responsible for more of the variance in the additive genetic factor.

Our analysis in Appendix III is more general and allows for assortative mating. While the resulting equation for α_g itself is more complicated, the conclusions about its interpretation remain valid. In contrast to α_g, Equation (11) implies that α_p does *not* generally have a causal interpretation, since $\mathbf{b}_p$ does not. Furthermore, in the general case with assortative mating, we show that α_p is a function of both $\boldsymbol{\beta}$ and $\mathbf{b}_p$. Because it can partly (or even wholly) reflect $\boldsymbol{\beta}$, it is wrong to interpret α_p as the "non-genetic" or "environmental" effect.[21]

Just as $\boldsymbol{\beta}$ is identified if regression Equation (8) includes the father's and mother's genotypes separately rather than the mean parental genotype, α_g is identified if regression Equation (9) includes the father's and mother's PGIs separately rather than the sum of parental PGIs. Including the father's and mother's PGIs separately enables comparisons of the magnitudes of the father's and mother's coefficients. More generally, α_g remains identified when controls are added to regression Equation (9), as long as the controls are not themselves caused by genotypes, because the random component of the child's PGI is independent of such controls. For including the father's and mother's PGIs separately and more generally for including controls, there are opposing effects on precision: adding covariates adds degrees of freedom but can absorb more of the residual variation.

Currently, rather than controlling for parental PGIs, it is more common to analyze sibling samples and control for family fixed effects. As in the case of estimating the effects of genotypes discussed in Section III.E, it is not widely appreciated that the regression with family fixed effects generates a biased estimator of the self genetic effect in the presence of sibling genetic effects and is inefficient relative controlling for the sum of parental PGIs. In the case of PGIs, the identifying variation with family fixed effects is the individual's PGI relative to the sibling mean, written here for the case of sibling pairs: $g_{\mathbf{w}} - (g_{\mathbf{w}} + g_{\mathbf{w},sib})/2 = (g_{\mathbf{w}} - g_{\mathbf{w},sib})/2 = (\mathbf{x} - \mathbf{x}_{sib})\,\mathbf{w}/2$, where the subscript "sib" denotes an individual's sibling. Variation in $(g_{\mathbf{w}} - g_{\mathbf{w},sib})/2$ is random, but when y is regressed on $g_{\mathbf{w}}$ controlling for family fixed effects, the coefficient that is estimated is

$$\alpha_g = \frac{\mathbf{w}'\boldsymbol{\Sigma}\,(\boldsymbol{\beta} - \boldsymbol{\beta}_{sib})}{\mathbf{w}'\boldsymbol{\Sigma}\mathbf{w}},$$

where $\boldsymbol{\beta}_{sib}$ denotes the sibling genetic effect. Since the identifying variation is the individual's PGI relative to her sibling's, the coefficient is picking up both the effect of the individual having a higher PGI and the effect of the sibling having a

[21] Under stronger assumptions, including that assortative mating is at equilibrium—meaning that the correlations between alleles do not change between generations (a common assumption in the literature that we do not make here)—Young (2023) proves a related result that provides some intuition. Specifically, Young expresses the coefficient on the parental PGI in terms of variances due to own and parental genetic effects under random mating and the correlations between these components within and across parents at assortative-mating equilibrium. Under Young's assumptions, his result shows that, unless the PGI captures all of the heritability, the parental PGI coefficient will include a contribution from own genetic effects because of LD induced by assortative mating. Young uses this result to propose estimators of heritability and variance due to parental genetic effects that adjust for the impact of assortative mating, but the estimators may be biased when his assumptions are violated.

lower PGI.

Analogous to the discussion in Section III.D, when sibling genotypes are observed but parental genotypes are not, Young et al. (2022) shows that controlling for family fixed effects is dominated by imputing parental genotypes and controlling for the sum of parental PGIs (constructed from the imputed data), with random effects to control for family-specific means. This strategy generates a consistent and unbiased estimator of Equation (10) with greater precision than the family-fixed-effects specification.

To date, most applications involving PGIs have estimated a regression like Equation (9) but with imperfect controls $\mathbf{z}$, such as principal components of the genetic data, instead of controlling for the sum of parental PGIs. Such a regression can be understood as estimating the parameter in Equation (10), but with omitted-variables bias due to uncorrected-for gene-environment correlation, as well as bias due to assortative mating.

E Correcting for PGI Measurement Error in Applications

In most social-science applications with PGIs, researchers use the PGI as a proxy for the total genetic influence on a phenotype. Indeed, as we showed in Section IV.A, the PGI can be interpreted as a standardized, noisy measure of the optimal predictor, where the measurement error is (approximately) classical. Unfortunately, errors-in-variables bias due to measurement error can distort empirical conclusions in a number of ways (e.g., Gillen, Snowberg and Yariv, 2019). Moreover, since PGIs vary by GWAS source and are constructed using different methods, the amount of measurement error varies. Both to reduce bias and to facilitate comparability across studies, it is useful to correct for the errors-in-variables bias in applications.

Two approaches have been developed for such corrections. First, DiPrete, Burik and Koellinger (2018) proposed an instrumental-variables approach. Two independent subsamples of the GWAS are used to construct two sets of PGI weights. These are then used to construct two PGIs in the prediction sample, and they are used to instrument for each other (as in Gillen, Snowberg and Yariv, 2019).

The other approach is a regression-disattenuation estimator, which uses external information on the amount of measurement error (Becker et al., 2021). The amount of measurement error can be estimated as the ratio of the optimal predictive power in the GWAS sample, $\breve{h}^2$, to the PGI's predictive power, R^2. The optimal predictive power is estimable using any of the genomic-relatedness or GWAS-based discussed in Section III.G, with controls $\mathbf{z}$ that are the same as in the GWAS. Ideally, R^2 should be estimated directly in a holdout subsample of the GWAS sample.

The instrumental-variables estimator has the advantage that it does not require an estimate of the optimal predictive power. This advantage may be particularly relevant for phenotypes with substantial assortative mating, which biases

estimates of optimal predictive power. The estimator's main drawback is a (potentially very substantial) loss of statistical power from having to split the GWAS sample. van Kippersluis et al. (2023) provides a detailed analysis.[22]

In this section, we have focused on adjusting for measurement error in the PGI to approximate the results if the *optimal predictor* had been analyzed. However, researchers often interpret PGIs as unbiased estimates of the *additive SNP factor*, an interpretation that is valid only if the GWAS controls are **sufficient for a credible causal interpretation**.[23] If we grant this premise, a natural question arises: can we further adjust our estimates to approximate what the results would look like if the additive genetic factor—or even the full genetic factor—had been analyzed instead? This would generally align better with theoretical models that motivate applications. Such a correction is indeed feasible. For example, the regression-disattenuation method could be applied using a twin-based estimate of heritability instead of SNP heritability. However, such a correction would rest on the untestable assumption that unmeasured variants that are **imperfectly correlated with** SNPs included in the PGI have the same relative effect magnitudes as the variants captured by the PGI, with the same relative magnitudes. This assumption is unlikely to hold because the unmeasured variants that **have low correlation with** SNPs are primarily rare variants, which may operate through different mechanisms than common SNPs (in particular, rare variants can have large biological effects). While applying such a correction might still be of interest, it is essential to be transparent about this additional assumption.

F PGIs As Social-Science Variables

Since a PGI is, at best, a proxy for the additive SNP factor, and the additive SNP factor is a proxy for the additive genetic factor, all of the interpretational caveats from Section II.E apply to PGIs. In particular, the effects of a PGI will typically reflect a mix of all the mechanisms through which genetic variants (that are correlated with included SNPs) operate. For many phenotypes, these mechanisms will include endogenous social and behavioral responses to phenotypes proximally affected by the PGI. Just as heritabilities are not measures of innateness, it is a mistake to assume that PGIs exclusively capture purely biological or innate characteristics.

Some researchers, especially non-economists, have asserted that it is misleading to describe the effects of a PGI as "causal" because the mechanisms are largely

[22] Corrections for errors-in-variables bias require some additional assumptions when multiple PGIs are included **in the regression**. Sanz-de Galdeano and Terskaya (2025) extend the Becker et al. approach to **regressions that** control for parental and sibling PGIs to estimate causal effects. Doing so requires assumptions about the parent-child or sibling correlation of the optimal predictor. Under random mating, these parameters are known and equal to $1/\sqrt{2}$ and $1/2$, respectively (see Trejo and Domingue, 2018), but more generally the parameters depend on the degree of assortative mating.

[23] If the controls in a population-based GWAS are not sufficient to eliminate bias, then there is an additional problem: the measurement error may not be classical because biases in population-based GWAS estimates may be correlated with the true causal effects.

or entirely unknown. Economists are well situated to provide a useful perspective, since economists often study the causal effects of environmental factors (and interventions) for which we have only a partial understanding of mechanisms.

As a variable that operates through many mechanisms, a PGI is like many other variables that social scientists study and incorporate into their theories. For example, an individual's biological sex has biological effects, such as body size and hormone levels, but it also affects an individual's behavior and outcomes through the reactions that other people have to the individual. While researchers need to bear these different possible mechanisms in mind when studying biological sex, it is nonetheless an important and useful variable in social-science research. We believe PGIs can be important and useful in a similar way.

Of course, for many purposes, it is important to understand the mechanisms underlying a causal effect. That is why, after credibly identifying a causal effect, many economics papers go on to study potential mechanisms—albeit often with evidence that is less air-tight than the identification of the causal effect. Carvalho (2025) offers a template for such analyses, in the context of studying *why* the PGI for educational attainment has a causal effect on educational attainment. Using a structural model, he finds evidence that the PGI reduces the marginal cost and increases the marginal benefit of education. In reduced-form analyses, he finds a positive causal effect of the PGI on the return to schooling and suggestive evidence that fluid intelligence and self-control (but not other personality traits) partly mediate the effect of the PGI on education.

V Applications

In this section, we critically discuss applications in economics that make use of genetic data. **While economists began working with genetic data earlier (e.g,, Fletcher and Lehrer, 2009), most of the recent work has analyzed PGIs. Pioneering contributions include Papageorge and Thom (2020) and Barth, Papageorge and Thom (2020). Papageorge and Thom (2020) addresses the longstanding question of how ability and human capital investments affect earnings. This literature has generally used cognitive test scores as an empirical proxy for ability, but cognitive test scores are affected by investments that have already been made before the tests are taken. Papageorge and Thom introduces the idea of instead using a PGI for educational attainment, which is fixed at conception and therefore predetermined with respect to any human capital investments. The paper reports three sets of main results. First, the association between the PGI and the probability of completing high school is decreasing in childhood socioeconomic status, but the association between the PGI and the probability of completing college is increasing in childhood socioeconomic status. Papageorge and Thom interpret this finding as suggesting that genes are complementary with family resources for college completion but substitutable**

for high school completion. Second, the PGI is strongly associated with wages for those who have completed college but not for those with less education. Combined with the first result, this finding implies that individuals with high PGIs are less likely have high earnings if they were born into disadvantaged circumstances. However, if cognitive test scores are used as the proxy for ability, then the relationship with wages is equally strong regardless of the level of education, and this implication would be obscured. In the third set of results, the PGI is used to control for genetic differences when examining the effects of specific aspects of childhood environment, such as physical abuse during childhood and drug or alcohol problems of parents. The paper estimates that such adverse circumstances have strong negative effects on later-life earnings and wealth.

Barth, Papageorge and Thom (2020) is motivated as an investigation into the persistence of the large and growing amount of wealth inequality in the U.S. Research suggests that one cause of wealth inequality is persistent differences in asset returns across individuals (e.g., Bach, Calvet and Sodini, 2020). Genetic effects on financial decision-making could contribute to the persistence of wealth inequality because genotypes are transmitted intergenerationally. Barth et al. find that a PGI for educational attainment predicts household wealth: a one-standard-deviation increase in the PGI is associated with a 23 percent increase in wealth. This relationship remains economically large and statistically significant after controlling for a number of factors, including education, income, and business ownership. This evidence suggests that the PGI is related to financial decision-making, and Barth et al. indeed find that the PGI is associated with stock market participation, financial literacy, planning horizons, and sophistication in probabilistic thinking. To further test if the genetic effects are mediated by financial decision-making, Barth et al. examine whether the PGI predicts wealth among households with defined benefit pensions. These are households that have very limited scope for financial decision-making. As predicted, Barth et al. find that the PGI does not predict wealth among these households. As papers we discuss below will illustrate, some of the subsequent literature has taken up the same research topics as these early papers and improved on them along several dimensions—most notably by more credibly identifying causal genetic effects—but these early papers played an important role in demonstrating the relevance of genetic data for core economic questions.

To facilitate the exposition in the remainder of this section, we organize the papers we discuss into categories that fall roughly along a continuum of conceptual complexity. We begin with applications where genetic variables are used solely for their predictive power, and conceptual issues are minimal. We end with appli-

cations that embed genetic variables into structural models or policy evaluations, where the identification and interpretive challenges are more demanding. In each case, we highlight research opportunities: in some cases a lack of existing work, and in other cases limitations of the work that has been conducted to date and how future work could improve on it. **Rather than trying to be exhaustive in our selection of papers, our aim is to convey the central issues for each type of application. For this reason, we discuss a smaller number of papers and emphasize the relevant details of each paper we discuss.**

A Polygenic Indexes for Balance Tests and as Covariates

The most conceptually straightforward use of genetic data in economics is as covariates or as balance-test variables in experimental and quasi-experimental designs **that estimate the effect of a treatment**. What matters here is only that the genetic variable—typically a polygenic index (PGI)—is predictive, not why it is predictive.

We highlight three reasons why PGIs are well suited for balance tests in **randomized controlled trials (RCTs) and** quasi-experiments. First, PGIs are predetermined characteristics, much like age or sex, so PGIs cannot be affected by the treatment. Second, the cost of genotyping participants may be small relative to the cost of collecting some alternative measures used in balance tests (e.g., scores from long cognitive tests). Finally, once participants have been genotyped, it is possible to construct PGIs for multiple phenotypes and genetic principal components, which can all be used for (in some cases, uncorrelated) balance tests. There are now a number of examples in the literature (e.g., Barcellos et al. 2018; 2025, and Schmitz and Conley, 2017); for example, Barcellos, Carvalho and Turley (2018) (their Appendix B) used PGIs for educational attainment and BMI, as well as 15 genetic principal components, as variables for a balance test for a regression-discontinuity design.

For the same reasons, PGIs may also be valuable control variables. For example, a PGI could be used to (partly) control for omitted variable bias in observational studies where the treatment of interest is correlated with genetic factors, e.g., studies of the association between parental behaviors and children's outcomes (e.g., Jami et al., 2021; for an alternative approach, see Zhao et al., 2025). Alternatively, **even in RCTs that** yield unbiased treatment effects without any control variables by virtue of randomizing the treatment, PGIs can be useful as controls that absorb residual variance, thereby making the treatment effect estimates more precise (Rietveld et al., 2013; Benjamin et al., 2012; Cesarini and Visscher, 2017).[24]

Rietveld et al. (2013) calculated the gains in effective sample size that could

[24]Controlling for PGIs can similarly increase the power of GWASs (see Bennett et al., 2021, Campos et al., 2023, and Jurgens et al., 2023). Relatedly, PGIs can be used for stratified sampling in an RCT, selecting extreme individuals to increase power for a given sample size (Fahed, Philippakis and Khera, 2022).

be obtained by controlling for PGIs in a simple RCT with two conditions (see Appendix IV for details). For example, if the set of baseline controls, absent the PGI, explain 20% of the variance in the outcome, they find that adding a PGI with an incremental R^2 of 15% would increase power equivalent to increasing the RCT sample size by 19%.

To date, only a handful of studies have used PGIs as control variables (e.g., Barcellos, Carvalho and Turley, 2018; Davies et al., 2018), perhaps due to lack of human capital for incorporating genetic-data collection into existing RCT research infrastructures. Some investigators may also be apprehensive about collecting sensitive data that is unrelated to the goals of the relevant RCTs. However, the benefit-cost ratio of controlling for PGIs will only grow as genotyping becomes cheaper and PGIs become more predictive and available for more phenotypes.

B Genetic Treatment-Effect Heterogeneity

Another category of applications examines heterogeneous treatment effects by genotype. In many contexts, it is useful to ask whether a policy or intervention has systematically different effects across individuals with different characteristics. As Manski (2011) notes, interacting treatments with genetic variables raises no special conceptual issues beyond those that arise when interacting with other predetermined characteristics. Such analyses can yield relevant insights even when the source of the heterogeneity is not fully understood. For example, they may serve as a source of hypotheses about underlying mechanisms, help identify subpopulations who are likely to benefit from a particular intervention, and motivate further work to understand the structure of treatment-effect heterogeneity. We believe that genetic treatment-effect heterogeneity is the second most common type of application of genetic data in the social sciences (Ahlskog et al., 2024; Biroli et al., 2025; Herd et al., 2019; Schmitz and Conley, 2017; Wedow et al., 2018), the most common being Mendelian randomization studies (discussed in Section V.F). **We illustrate this category of applications and the limitations of research to date by briefly summarizing a few recent studies that offer** useful templates for how such analyses can inform both theory and policy. **For excellent and more in-depth discussions, see Biroli et al. (2025) and Miao et al. (2025).**

Barcellos, Carvalho and Turley (2018) estimate how the treatment effects of education on health vary by PGI using a regression-discontinuity design that, following Clark and Royer (2013), exploits a British schooling reform from 1972 that raised the compulsory schooling age from 15 to 16. Using data from UK Biobank, they find that, for example, an additional year of schooling reduced the risk of obesity by only 0.3 percentage points for those with a BMI PGI one standard deviation below the mean but reduced it by 11.7 percentage points for those one standard deviation above. Basu et al. (2025) examine whether PGIs for smoking behaviors moderate the effectiveness of a randomized smoking-cessation intervention in the Lung Health Study (Anthonisen et al., 1994) that combined

behavioral counseling, nicotine gum, and, in some arms, pharmacological support. On average, individuals in the treatment group were 23 percentage points more likely to quit smoking than those in the control group. However, a one-standard-deviation increase in a PGI for smoking initiation was associated with a 2.5-percentage-point reduction in the probability of cessation.

These studies share three limitations with nearly all such work to date. First, although the policy variable is exogenous, the genetic variable is not. **The framework discussed in this paper clarifies the value of controlling for parental PGIs to credibly identify causal genetic effects; extending this reasoning to the context of treatment-effect heterogeneity, credible identification of the interaction of the treatment with the PGI requires controlling for both the parental PGIs and the interaction between the parental PGIs and the treatment.** Second, neither paper makes much progress on elucidating mechanisms. **Third, the analyses do not correct for measurement error in the PGIs (see Section IV.E above). Thus, they probably substantially underestimate the true magnitude of the heterogeneity by genotype.**

A recent paper, Biroli et al. (2025), overcomes the first of these limitations. Analyzing data from the Avon Longitudinal Study of Parents and Children, the paper exploits the UK's strict birth-date cut-offs for school entry to isolate quasi-random variation in whether individuals are old for their grade. Consistent with prior work, the paper finds positive effects of being old-for-grade on standardized test scores. Biroli et al.'s main question is how this treatment effect varies with the PGI for educational attainment. To isolate exogenous variation in the PGI, Biroli et al. control for (measured or imputed) mean parental PGI in the analysis and its interaction with the treatment variable. They find that individuals with higher PGIs benefit more from being old-for-grade for the Entry Assessment test (taken before the start of schooling) but benefit less from being old-for-grade for test scores later in childhood.

Another limitation is inherent to all gene-by-environment interaction studies that use PGIs: they restrict attention to environmental interactions with a fixed linear combination of SNPs (see e.g., Tahmasbi et al., 2017). We anticipate that future work will increasingly use alternative methods that relax this restriction. Wang et al. (2019) demonstrate one way to do so: they conduct a GWAS where the **dependent variable is essentially** the *variance* of a phenotype, rather than its level. When a SNP's genotype is related to variance in the phenotype, the presence of heterogeneous genetic effects can be inferred (see also Johnson, Sotoudeh and Conley, 2022).

Miao et al. (2025) **recently developed a powerful framework for conceptualizing and estimating gene-environment environments without using PGIs. Builds on the assumptions and approach of LD Score regression**

(see Section III.G), they show how the variance in a phenotype attributable to gene-by-environment interactions can be estimated from the summary statistics of a ***genome-wide interaction study (GWIS)*: a GWAS whose regression specification includes SNP-by-environment interactions. Their estimator is not subject to the errors-in-variables bias that arises from measurement error in a PGI. Maio et al. also show, under their assumptions, that the coefficient-based genetic correlation between the SNP-by-environment interaction effects from the GWIS and the SNP effects from a GWAS of some phenotype is equivalent to the interaction effect between the environmental variable and the additive SNP factor of the phenotype. They illustrate their approach by extending and replicating the analysis of Barcellos, Carvalho and Turley (2018) dicussed above. In the same UK Biobank data and using the same regression-discontinuity design as Barcellos et al., they run a GWIS by regressing a measure of health in middle age on each SNP, years of schooling, and the SNP-by-years-of-schooling interaction. Using the same GWAS summary statistics that Barcellos et al. used to construct a PGI for educational attainment, Maio et al. apply LD Score regression to estimate the coefficient-based genetic correlation between the GWIS and GWAS summary statistics. Their results match those of Barcellos et al. after correcting the latter for measurement error in the PGI. We anticipate that approaches like Maio et al.'s, which build on methods from statistical genetics to address questions of interest in the social sciences, will become increasingly influential in the coming years.**

C Policy Analysis of Genetic Advances

The accuracy of genetic predictions for disease, mortality, and other phenotypes will continue to improve in the coming years, raising a host of complex policy and regulatory questions about their effects on various markets. Economists are well equipped to contribute to analyzing these issues.

For example, Azevedo, Beauchamp and Linnér (2024) examine the potential impact of improved PGI accuracy on critical illness insurance markets, which pay out a lump sum upon the diagnosis of any covered condition. The conceptual issues here related to genetics are relatively simple, related primarily to the incremental predictive power of future PGIs (beyond the currently available information). To forecast this, **Azevedo et al.** begin with the observation that as the volume of available training data increases, the predictive accuracy of future PGIs will approach **the** asymptotic-R^2 term in Equation (7). Using UK Biobank data, they study the effects of widespread access to such PGIs on future insurance markets, finding that it is likely to generate alarmingly high levels of adverse selection that could lead to unsustainably high premiums or even market unraveling.

As another example, PGIs have begun to be integrated into healthcare systems

with the goal of enabling more effective, targeted treatments, raising questions about whether and how PGIs should be incorporated into screening (e.g., Schunkert et al., 2025). In this context, it is not sufficient to accurately assess future predictive power, since the ability to identify at-risk individuals with high accuracy is only valuable if there is an intervention that passes a cost-benefit test in the identified group.

To give one more example, in the United States, some companies now offer couples undergoing in vitro fertilization the option to screen embryos for polygenic phenotypes, in addition to monogenic diseases (Roura-Monllor et al., 2025). While this technology has the potential to reduce population morbidity and healthcare costs, it can also generate externalities (e.g., if couples select for taller children but relative height is what matters) and create an additional channel for intergenerational transmission of inequality. In these and other cases, an economic framework can contribute to optimal policy design.

D Assortative Mating

Assortative mating is of particular interest to economists because matches are equilibrium outcomes shaped by preferences, constraints, and strategic considerations, and understanding these processes is crucial for analyzing inequality, intergenerational mobility, and household resource allocation. Genetic data has three properties that make it a valuable new tool for studying assortment processes.

First, genotypes—and hence a spouse's PGI—**are** fixed at conception. Thus, as noted by both Conley et al. (2016) and Robinson et al. (2017), a positive spousal genetic correlation for a phenotype must reflect factors in place prior to the match, whereas a phenotypic correlation need not. Consider, for example, BMI. A positive spousal phenotypic correlation could reflect correlated spousal environments after marriage (e.g., due to shared meals and habits). In a sample of 24,662 spousal pairs (from the UK Biobank, 23andMe, and several other datasets), Robinson et al. (2017) confirm that the **measurement-error-corrected** coefficient from a regression of a spouse's BMI on their partner's PGI for BMI (0.143, SE = 0.007) is indeed lower than the phenotypic correlation (0.228, SE = 0.004). As a placebo test, they report the same analyses for height, a phenotype that is largely fixed prior to the match, and find that the two estimates are indistinguishable (0.200, SE = 0.004, versus 0.201, SE = 0.004, respectively).

Second, since genotypes are transmitted to offspring, sorting on genotypes generates a mechanism for persistence of inequality across generations. Abdellaoui et al. (2022*b*) formalize this insight, building an economic model in which a person's socioeconomic status (SES) and their advantageous genes are both assets in the marriage market on which partners sort. Advantageous genes and SES are both transmitted and thus become correlated in subsequent generations. To test this model, using the UK Biobank, **Abdellaoui et al.** measure whether later-born children—who have lower SES on average than their older siblings—tend to

marry people with lower average educational-attainment PGIs. They find weak evidence that later birth (and therefore lower expected SES) is associated with a person's spouse having a smaller educational-attainment PGI (-0.031, SE: 0.015 in their strongest specification), though this result is not robust across all specifications.

Third, genetic data on a cross-section of individuals can be used to make inferences about assortment in previous generations even without data on spouse pairs and even without phenotype data! Such inferences are possible because if parents (or earlier ancestors) sort on a heritable phenotype, then alleles that **cause** increases in the phenotype will be correlated across the father and mother and thus correlated *within* their child's genome, even if the alleles are on different chromosomes. Yengo et al. (2018) made this observation and exploited it to construct an estimator for the amount of assortative mating in some phenotype: they infer it from the correlation between a PGI that is constructed only from SNPs on even-numbered chromosomes and another constructed only from SNPs on odd-numbered chromosomes. **Yengo et al.** find positive cross-chromosome correlation for height and educational attainment, consistent with evidence for assortative mating on these phenotypes based on observed spouse pairs.

Much of the work **on assortative mating** to date shares two limitations. First, like other applications, it uses PGIs from population-based GWASs. Consequently, the estimates of genotypic sorting may be confounded by any gene-environment correlation that the GWAS did not fully control for. As PGIs from sufficiently well-powered family-based GWASs become available, applications should use those instead. Second, the work implicitly assumes that mates' genotypic correlation and mates' phenotypic correlation are directly comparable. In fact, however, their dynamics differ (see, e.g., Crow and Kimura, 1970). For example, a one-time, permanent increase in the amount of phenotypic assortment on height would generate a gradual increase in the correlation between mates' PGIs for height over several generations, asymptoting toward a higher level. Future research should account for these dynamics when interpreting results.

E Interpersonal Genetic Effects

As discussed in Section II.F, interpersonal genetic effects—the effect of one person's genome on someone else through influencing the other person's environment—can provide a valuable new source of evidence for studying how people affect each other. In the past few years, a burgeoning literature has estimated friend, parental, and sibling genetic effects using PGIs. Domingue et al. (2018) examined the association of school friends' PGIs for height, BMI, and educational attainment with one's own phenotype values and found a positive relationship for educational attainment. Following Kong et al., 2018, many papers have estimated the association of parents' PGIs for educational attainment with their children's phenotypes (e.g. Armstrong-Carter et al., 2020). Domingue and Fletcher (2020) contrast these associations for biological versus adopted children. A few papers

have studied the association between a sibling's PGI and one's own phenotype, including Cawley et al. (2019), who find a positive association for obesity, and Cawley et al. (2023), who report a null result for educational attainment (Howe et al., 2022*a*, explore an alternative approach based on singletons). Unfortunately, none of these results have a clean causal interpretation because the associations could be driven by uncorrected-for gene-environment correlation: for example, in the case of estimating sibling genetic effects, a sibling's PGI is correlated with the parents' PGIs, which affect the family environment.

The best-identified work to date we are aware of is Young et al. (2022), who estimated sibling genetic effects by regressing the individual's phenotype on their own PGI, their sibling's PGI, and the (measured or imputed) parental PGIs. Because the sibling's PGI is random conditional on the parental PGIs, this design has a causal interpretation. In UK Biobank data, Young et al. find no evidence for sibling genetic effects of the educational attainment PGI on a range of phenotypes, including educational attainment, cognitive ability, BMI, and smoking.

Future work on sibling and friend genetic effects should similarly control for parental PGIs. Cleanly identifying parental genetic effects turns out to be trickier because—even if controlling for grandparental PGIs so that parental PGIs are conditionally random—the coefficient on the parental PGIs is biased in the presence of assortative mating. Consider assortative mating on educational attainment. The random component of the father's PGI is correlated with father's education, which is correlated with mother's education, which is correlated with the mother's additive genetic factor for education—including the component that is not captured by the mother's PGI. **Half of this** component is transmitted to the child and can have a self genetic effect on the child (the same issue biases the study design in Nivard et al., 2024). To overcome this limitation, studies of parental genetic effects will need to model and correct for this assortative-mating effect, in addition to controlling for grandparental PGIs.

F Mendelian Randomization

There is a vast epidemiological literature that uses genetic variants as **instrumental variables** to make causal inferences about the effects of various "exposures" on health and behavioral outcomes—a strategy known as Mendelian randomization (MR) (Burgess et al., 2020*b*; Davey Smith and Ebrahim, 2003). The earliest applications of genetic data in economics were MR studies (Norton and Han, 2008; Ding et al., 2009; Fletcher and Lehrer, 2009; von Hinke Kessler Scholder et al., 2011). For skeptical reactions, see, e.g., Beauchamp et al. (2011), Benjamin et al. (2012), Conley (2009), Cawley, Han and Norton (2011), and McMartin and Conley (2020). The number of MR studies relevant to economics likely exceeds that of all other applications of genetic data combined.

Like any instrumental variables analysis, an MR study's credibility typically hinges on how plausibly the exclusion restriction can be defended—that is, whether it can be credibly established that the genetic variant influences the

outcome solely through its effect on the exposure. We emphasize three factors that merit close attention when evaluating this identifying assumption (for a more extensive discussion and primer on MR, see Davey Smith and Hemani, 2014, and von Hinke et al., 2016). First, the assumption is more plausible when **the variant is known to cause the exposure (either through credible causal identification of the variant's effect or through understanding the biological pathway linking the variant to the exposure).** Second, it is more defensible when researchers can rule out alternative pathways or correlations with other determinants of the outcome—for example, those arising from population stratification or from causal effects of a given genetic variant on more than one phenotype (often called *horizontal pleiotropy*[25]). Third, results from placebo tests**, sometimes called "negative controls,"** conducted in subpopulations where the exposure is **absent** can provide some reassurance. Failure to detect a reduced-form relationship between the variant and the outcome in such populations can help mitigate concerns about alternative causal pathways.

When all three conditions are met, MR offers a compelling strategy for strengthening causal inference beyond what can be achieved through conventional analyses of observational data. One example of a persuasive MR study is Millwood et al. (2019), which investigates the effect of alcohol consumption on cardiovascular disease outcomes and blood pressure, with particular attention to stroke. Prior observational studies had reported a J-shaped relationship between alcohol consumption and health: **while heavy drinkers were the least healthy, moderate drinkers were healthier** than abstainers, causing some researchers to hypothesize that light drinking might be protective. However, this pattern could reflect confounding or reverse causation—for instance, if individuals in poor health are more likely to abstain or if moderate drinking is associated with other unobserved factors conducive to good health. Millwood et al. (2019) use genetic variants in the genes *ALDH2* and *ADH1B* as instruments for alcohol consumption in a sample of approximately 160,000 Han Chinese adults from the China Kadoorie Biobank. Both of the instruments are well understood biologically: **they both affect the speed of alcohol metabolism, and individuals with slow metabolism experience unpleasant reactions to alcohol and thus consume less.** The study's IV estimates suggest that increased alcohol consumption raises systolic blood pressure and increases the risk of stroke regardless of the baseline level of consumption. The monotonic relationship is at odds with the hypothesis that moderate levels confer health benefits. To probe the exclusion restriction, the authors conduct a placebo test in women, most of whom abstain from alcohol for cultural reasons. In this subsample, they find no association

[25]Loosely speaking, a genetic variant is said to be pleiotropic when it affects more than one phenotype, but not all types of pleiotropy are problematic for MR. "Horizontal pleiotropy" refers to when a genetic variant affects the phenotypes through distinct biological mechanisms, or when the phenotypes are causally unrelated to each other. "Vertical pleiotropy" refers to when a genetic variant affects a phenotype that in turn affects another phenotype. While horizontal pleiotropy may lead to violation of the exclusion restriction, the exclusion restriction is a case of vertical pleiotropy.

between the genetic variants and cardiovascular outcomes, supporting the interpretation that the variant affects outcomes only through alcohol consumption.

We view studies like Millwood et al. (2019) as outliers in terms of credibility relative to the literature as a whole. Advocates of MR contend that, despite its limitations, it can provide more credible and externally valid evidence than alternative study designs, especially when randomized experiments are infeasible; for example, for studying the causal effect of childhood height (von Hinke Kessler Scholder et al., 2013) or estimating the marginal healthcare costs associated with health conditions (Dixon et al., 2016). Moreover, to address concerns about both violations of the exclusion restriction and weak instruments, a growing methodological literature has developed methods that, by using many genetic variants as instruments, rely on weaker identifying assumptions (e.g., Bowden, Davey Smith and Burgess, 2015; Brumpton et al., 2020; Burgess et al., 2020*a*). Advocates argue that credibility can be accumulated by "triangulation"(Munafò and Davey Smith, 2018): convergent evidence from estimators with different identifying assumptions. Burgess et al. (2020*b*; 2023) outline best practices for MR studies. Our own view is that, especially in social-science contexts, the identifying assumptions for MR are generally unlikely to hold. Nonetheless, MR studies can be valuable in settings where, relative to the biases from other feasible study designs, the bias from violations of the MR assumptions is likely to be small.

G Incorporating PGIs into Structural Models

Applications that explicitly incorporate PGIs into structural models allow for richer inferences about the mechanisms underlying the mapping from genotypes to outcomes under the maintained assumptions of the model. We discuss three examples that share some features in common. The first two papers build on prior work that relied on measures of a child's ability, such as cognitive test scores, that are themselves influenced by parental investments. **Following the pioneering work of Papageorge and Thom (2020),** the papers instead use a PGI for educational attainment, which has two key advantages: it is fixed at conception, prior to any (even in-utero) parental investments, and it is randomly assigned, conditional on parental PGIs. All three papers use genotyped family data and control for parental PGIs in their analysis, with the first two applying the method developed by Young et al. (2022) (see Section III.D) to impute missing mean parental genotypes.

The first paper is Sanz-de Galdeano and Terskaya (2025), which addresses the question of whether parental investments compensate for or reinforce children's ability differences. They use data from 604 genotyped sibling pairs with European genetic ancestries in the National Longitudinal Adolescent to Adult Health Study. Based on a survey conducted when the children were aged 12-20 that asked about how often the child engaged in various activities with the parents, Sanz-de Galdeano and Terskaya construct an index of parental investment for

each child. The paper's main regression specification is

$$(12) \qquad I_0 = \beta_0 + \beta_1 \left(\text{PGI}_0 - \text{PGI}_y\right) + \beta_2 \text{PGI}_0 + \beta_3 \text{PGI}_{par} + \text{Controls} + u,$$

where the unit of analysis is the sibling pair, I_0 is the index of parental investment in the older sibling, PGI_0 and PGI_y are the older and younger sibling's respective PGIs, PGI_{par} is the (imputed) parental mean PGI, and u is an error term. Because PGI_0 and $(\text{PGI}_0 - \text{PGI}_y)$ are conditionally random given PGI_{par}, the coefficients β_1 and β_2 have causal interpretations. In a structural model that extends Becker and Tomes (1976) and Behrman, Pollak and Taubman (1982), Sanz-de Galdeano and Terskaya show that β_1 captures a parental preference parameter for inequality aversion versus efficiency, and β_2 captures the cost of investment (the net effect of the child's PGI on parental investment is $\beta_1 + \beta_2$). As discussed in Section IV.E, estimating regression Equation (12) with the observed PGIs would generate coefficients that suffer from substantial errors-in-variables bias. Instead, Sanz-de Galdeano and Terskaya develop and apply an extension of Becker et al.'s (2021) measurement-error correction, which simultaneously corrects for the measurement error in all three PGI terms.

In their full-sample analysis including all their control variables and after correcting for the measurement errors in the PGIs, Sanz-de Galdeano and Terskaya estimate $\hat{\beta}_1 = -0.207$ (S.E. $= 0.102$), $\hat{\beta}_2 = 0.167$ (S.E. $= 0.140$), and $\hat{\beta}_3 = -0.029$ (S.E. $= 0.161$). The main result is the negative estimate of the parameter β_1, which suggests parents are inequality-averse over their children's human capital. This result is statistically weak, but Sanz-de Galdeano and Terskaya show that **their** measurement-error correction generates standard errors that are biased upward. The point estimate of β_2 is positive, which would imply parents invest more when their children have a higher PGI (conditional on the sibling difference), but the 95% confidence interval is large and includes zero. The estimate of β_3 is difficult to interpret because it picks up the effects of non-genetic variables that are correlated with PGI_{par}.

The second paper, Houmark, Ronda and Rosholm (2024), models and estimates the joint evolution of cognitive skills and parental investments throughout early childhood (building on Cunha and Heckman, 2007, 2008). They use data from 4,510 genotyped children and their parents (with European genetic ancestries) in the Avon Longitudinal Study of Parents and Children, a birth cohort study based in Bristol, UK. Genetic data from both parents are available for 1,267 children. For other children, the missing parent's genotype is imputed. Based on questionnaires sent regularly to the child's primary caregiver starting prior to birth, the authors construct measures of children's skills (e.g., ability to process new information and learn abstract concepts) and parental investments (the frequency with which the parent does certain activities with the child).

Houmark et al.'s model has three structural equations: (i) a child's initial cognitive skills as a function of the child's, the mother's, and the father's additive

SNP factors for educational attainment, G_i, G_i^m, and G_i^f; (ii) the production function for cognitive skills in each period t, which depends on the three additive SNP factors as well as parental investment; and (iii) parental investment behavior in each period t as a function of the three additive SNP factors as well as the child's cognitive skills in period t. Following standard practice in this literature (e.g., Agostinelli and Wiswall, 2016; Cunha and Heckman, 2008), the structural equations are supplemented by a set of measurement equations that link the unobserved theoretical constructs (cognitive skills, parental investments, the additive SNP factor) to the observed measures, under standard but strong i.i.d. assumptions on the measurement errors of the observed measures. By estimating these equations jointly, they adjust for the measurement errors in the PGIs.

The main results are about the effects of children's genotypes. These estimated effects have a causal interpretation due to the conditional random assignment of G_i, given G_i^m and G_i^f. The paper finds that genetic influences affect cognitive skills even for very young children, ages 0-2, and that the genetic influence on a child's cognitive skills is increasing with age. Previous work also reported increasing genetic influences with age (e.g., Bouchard, 2013; Belsky et al., 2016), but an alternative interpretation of the earlier findings—ruled out here—was that cognitive skills at younger ages are measured with more error. The paper also finds that children with higher additive SNP factors behave in ways that cause their parents to invest more in them. The parental investment responses magnify initial differences between children.

The other estimated effects should be interpreted more cautiously because they rely more heavily on the assumptions of the structural and measurement models, but they paint a rich picture of the dynamics of parental investment and children's accumulation of cognitive skill. For example, the paper finds that parents with higher additive SNP factors invest more in their children (holding fixed the child's additive SNP factor) and that the returns to parental investments are substantially overestimated if genetic measures are omitted from the analysis.

The third paper, Rustichini et al. (2023), improves on models of intergenerational mobility, which usually assume an ad hoc, exogenous equation for intergenerational skill transmission. Instead, Rustichini et al. endogenize skill transmission, microfounding it with models of genetic inheritance and assortative mating **imported from the genetics literature**. Using data from the Minnesota Twin Family Study, they estimate an overlapping-generations model using a PGI for educational attainment as an empirical proxy for genetic factors influencing skill. Like Houmark et al., they adjust for the measurement error in their empirical proxies by jointly estimating measurement equations that make strong i.i.d. assumptions. Rustichini et al. conclude that the standard model is likely to underestimate the intergenerational elasticity of income and that cognitive skills, more than personality traits, mediate the genetic effects.

Relative to most other applications to date, these three papers are methodologically strong: their empirical work is closely tied to state-of-the-art economic

models, the PGI estimates have a causal interpretation because they control for parental PGIs, and they correct for measurement error in the PGI. However, all three likely understate the economic significance of children's genotypes because, relative to the theoretical concept of "initial ability" that the PGI proxies for, the PGI contains additional measurement error that is not accounted for. For example, the theoretical concept corresponds to a PGI constructed from causal-effect estimates, as could be generated from a family-based GWAS, rather than the PGI these papers use, which comes from a population-based GWAS (see Section III.F). The theoretical concept also would fully capture the effects of all genetic variants, not just those captured by SNPs included in the PGI (see Sections II.C and IV.D). Although these sources of measurement error are neither classical nor mean-zero, their primary effect on the main results is likely to be an attenuation bias (Trejo and Domingue, 2018). Future research should adjust for these sources of measurement error (**for example, Alemu et al., 2025*b* recently proposed a way of doing so**). Although adjusting for these measurement errors requires additional assumptions (see Section IV.E), doing so is well within the spirit of structural modeling.

VI Future Directions and Concluding Remarks

Over the last ten years, with the advent of GWAS for social and behavioral phenotypes, social-science genomics has come of age. PGIs are beginning to be used in social-science applications. In some cases, PGIs will be useful as control variables to increase statistical power (e.g., in randomized experiments) or to address confounds, for example, when studying the health-education gradent. In other cases, equipped with the PGI as a measure of genetic influences, economists and other social scientists will have greater leverage in addressing classic topics, such as the determinants and impacts of parental and school investments (e.g., Houmark, Ronda and Rosholm, 2024; Sanz-de Galdeano and Terskaya, 2025), labor market returns to human capital (e.g., Papageorge and Thom, 2020), intergenerational transmission of skills (e.g., Barth, Papageorge and Thom, 2020; Rustichini et al., 2023), the determinants and consequences of migration (e.g., Abdellaoui et al., 2019), and assortative matching in marriage markets (e.g., Abdellaoui et al., 2022*b*). Progress on these topics is already underway, as illustrated by some of the examples in Section V.

While some of these applications involve economists relatively straightforwardly importing PGIs from genetics into economics, in other applications, economists build structural models to account for endogenous behavioral and social responses to genotypes. In such applications, economists may contribute to geneticists' understanding of the mechanisms through which PGIs matter. Section V.G showcased some early examples.

To facilitate certain applications, we anticipate that social scientists will influence the genetics research that is conducted. This has already happened in the case of

GWASs for social and behavioral phenotypes, which have been collaborations between social scientists and geneticists, driven (at least initially) by social scientists' interests in the phenotypes. Once genotyped samples become large enough for adequate power, we anticipate that social scientists will want to conduct GWASs in samples that contain randomized experiments or quasi-experiments (see also Schmitz et al., 2021). For example, in a sample where the curriculum is randomly assigned, economists may be interested in a GWAS of educational attainment in which the regressors include SNP-by-treatment interactions. A PGI can then be constructed from the coefficients on the interactions. This PGI would capture genetic influences on the effectiveness of the treatment. When based on a sufficiently well powered GWAS, such a PGI would be a better tool for targeting the curricular intervention than the current PGI for educational attainment.

We anticipate that five ongoing developments will facilitate applications of genetic data in the social sciences. First is simply more and larger GWAS samples. **To date, the vast majority of applications have analyzed a PGI for educational attainment, largely because it has had the greatest predictive power among the PGIs relevant for social science. Accordingly, the applications have focused on research questions related to education and human capital. As large GWAS samples enable researchers to construct more highly predictive PGIs for other phenotypes, the range of possible applications will grow.** Moreover, larger samples will continue to enable better powered studies of all kinds, such as studies of gene-environment interactions. In addition, at some point, larger samples will enable researchers to construct PGIs that capture some of the non-additive genetic variance, allowing the predictive power of a PGI to exceed the phenotype's "optimal predictive power" (which is based only on additive effects). While we expect dominance variance to be negligible for most social and behavior phenotypes and epistatic variance to be small (see Section II.B), there is more uncertainty about epistatic variance. Although the combinatorial explosion of potential gene-gene interactions is a challenge for efforts to credibly identify the effect sizes of specific interactions, machine-learning methods should be able to capture some of their predictive power.

Second, as larger genotyped family samples become available, researchers will be able to more precisely estimate—**with credible causal identification**—self genetic effects, parental genetic effects, sibling genetic effects, and genetic effects of other family members. **As PGIs constructed from family-based GWASs become more predictive, researchers will be able to use them in applications in place of PGIs constructed from population GWASs. While a PGI from a population GWAS is a noisy measure of the optimal predictor, a PGI from a family-based GWAS is a noisy measure of the additive SNP factor (see Sections IV.A and IV.E). As such, the latter is typically closer to the theoretical construct of interest.**

Third, generating large samples with genetic data from all major underrepresented

genetic ancestries is a high priority. For the social sciences, **a major benefit is that it will be possible to construct** PGIs that are more predictive for individuals with non-European genetic ancestries.

Fourth, genotyping will become denser, implying that the measured SNPs will capture a greater fraction of all of the genetic variation. Indeed, as the cost of sequencing continues to plummet, sequencing may soon overtake **SNP-array** genotyping as researchers' preferred way of measuring genetic variation. Since sequencing can measure rare SNPs and non-SNP types of genetic variation much more accurately than **SNP-array** genotyping, it enables the discovery of very rare variants with large effects, e.g., on intellectual disabilities (e.g., Chen et al., 2023), that evade detection in association studies limited to (relatively common) SNPs. This will lead to a much better catalog of the genes that, when disrupted, have a large impact on phenotypes relevant to the social sciences. We expect that the main benefit of denser genotype measurement for social science will be the improved predictive power of PGIs. Moreover, in the limit where a PGI includes all genetic variants weighted by estimates of their causal effects, the PGI can be interpreted as a noisy measure of the additive genetic factor.

Fifth, genetic data will become increasingly available in datasets with rich behavioral data that are particularly valuable for social-science applications. These datasets already include the Health and Retirement Study, the Child Development Supplement of the Panel Study of Income Dynamics, and the German Socioeconomic Panel.

More so than in economics, progress in genetics has been propelled by technological advances in measurement that show no sign of slowing down. The nearly 15 years since the last time a review of "genoeconomics" has been written (Beauchamp et al., 2011; Fletcher, 2011; Benjamin et al., 2012) is an eternity in terms of the pace of genetics research. Although the coming 15 years may not produce transformational changes on par with those of the past decade and a half—a re-shaping of social-science genomics with GWAS and the resulting PGIs—there will surely be both progress and challenges that we cannot currently imagine.

We conclude by highlighting perhaps the most important ongoing challenge: conducting, interpreting, and communicating research at the intersection of genetics and social science responsibly. While these obligations apply to all researchers, researchers in social-science genomics bear additional responsibilities in light of how difficult it is to correctly interpret genetic associations—as highlighted by the extensive discussion of interpretation throughout this review—as well as the enduring legacy of eugenics (Rutherford, 2022). While far from sufficient, terminology can help to some degree. Researchers should be cognizant of the potential social harms of, and be especially careful about conducting and communicating, research that could be (mis)understood as comparing ethnic, racial, or other groups on socially valued phenotypes, such as cognitive performance or income. Given how easy it is to slip into genetic determinism, we believe it is helpful to continually remind readers of research papers that the effects of individual genetic

variants are small (e.g., Chabris et al., 2015), can operate through environmental pathways (Jencks, 1980), and have no obvious bearing on the effectiveness of interventions (Goldberger, 1979). We believe it is useful to write a Frequently Asked Questions (FAQs) document along with a paper to explain to journalists and non-experts what the research does and does not find and to carefully address any ethical or policy questions raised by the research. Indeed, writing such FAQs has become standard practice in social-science genomics (Martschenko et al., 2021). We recommend a report published by the Hastings Center, a bioethics think tank (Meyer et al., 2023), for helpful discussion of these and other best practices, as well as ethical issues related to social-science genomics more broadly.

REFERENCES

Abdellaoui, Abdel, Conor V. Dolan, Karin J. H. Verweij, and Michel G. Nivard. 2022*a*. "Gene–Environment Correlations Across Geographic Regions Affect Genome-wide Association Studies." *Nature Genetics*, 54(9): 1345–1354.

Abdellaoui, Abdel, David Hugh-Jones, Loïc Yengo, Kathryn E. Kemper, Michel G. Nivard, et al. 2019. "Genetic Correlates Of Social Stratification In Great Britain." *Nature Human Behaviour*, 3(12): 1332–1342.

Abdellaoui, Abdel, Oana Borcan, Pierre-Andre Chiappori, and David Hugh-Jones. 2022*b*. "Trading Social Status for Genetics in Marriage Markets: Evidence from UK Biobank." Human Capital and Economic Opportunity Working Group Working Paper No. 2022-018. Available at: https://hceconomics.uchicago.edu/research/working-paper/trading-social-status-genetics-marriage-markets-evidence-uk-biobank.

Agostinelli, Francesco, and Matthew Wiswall. 2016. "Estimating The Technology Of Children's Skill Formation." National Bureau of Economic Research Working Paper 22442. Available at: https://www.nber.org/system/files/working_papers/w22442/w22442.pdf.

Ahlskog, Rafael, Jonathan Beauchamp, Aysu Okbay, Sven Oskarsson, and Kevin Thom. 2024. "Testing for treatment effect heterogeneity: Educational reform, genetic endowments, and family background." *SSRN Working Paper*. Available at: https://ssrn.com/abstract=4758247.

Alemu, Robel, Alexander S. Young, Daniel J. Benjamin, Patrick Turley, and Aysu Okbay. 2025*a*. "Dissecting the Predictive Accuracy of Polygenic Indexes for Behavioral Phenotypes Across Genetic Ancestries." *bioRχiv*. Available at: https://www.biorxiv.org/content/10.1101/2025.09.11.675704v1.

Alemu, Robel, Anastasia Terskaya, Matthew Howell, Junming Guan, Harry Sands, et al. 2025*b*. "An Updated Polygenic Index Repository: Expanded Phenotypes, New Cohorts, and Improved Causal Inference." *bioRχiv*. Available at: https://pubmed.ncbi.nlm.nih.gov/40463245.

Anthonisen, Nicholas R., John E. Connett, James P. Kiley, Murray D. Altose, William C. Bailey, et al. 1994. "Effects of Smoking Intervention and the Use of an Inhaled Anticholinergic Bronchodilator on the Rate of Decline of FEV1: The Lung Health Study." *JAMA*, 272(19): 1497–1505.

Armstrong-Carter, Emma, Sam Trejo, Liam J. B. Hill, Kirsty L. Crossley, Dan Mason, and Benjamin W. Domingue. 2020. "The Earliest Origins of Genetic Nurture: The Prenatal Environment Mediates the Association Between Maternal Genetics and Child Development." *Psychological Science*, 31(7): 781–791.

Azevedo, Eduardo M., Jonathan Beauchamp, and Richard Karlsson Linnér. 2024. "Genetic Prediction and Adverse Selection." *The Wharton School Research Paper*. Available at: https://ssrn.com/abstract=5103439.

Bach, Laurent, Laurent E. Calvet, and Paolo Sodini. 2020. "Rich Pickings? Risk, Return, and Skill in Household Wealth." *American Economic Review*, 110(9): 2703–2747.

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 of the United States of America*, 115(42): E9765–E9772.

Barcellos, Silvia, Leandro Carvalho, Kenneth Langa, Sneha Nimmagadda, and Patrick Turley. 2025. "Education and Dementia Risk." *NBER Working Paper No. 33430*. Available at: https://www.nber.org/papers/w33430.

Barth, Daniel, Nicholas W. Papageorge, and Kevin Thom. 2020. "Genetic Endowments and Wealth Inequality." *Journal of Political Economy*, 128(4): 1474–1522.

Basu, Shubhashrita, Jason Fletcher, Qiongshi Lu, Jiacheng Miao, and Lauren Schmitz. 2025. "Understanding the Role of Genetic Heterogeneity in Smoking Interventions: Experimental Evidence from the Lung Health Study." *NBER Working Paper No. 33473*. Available at: https://www.nber.org/papers/w33473.

Bayarri, María Jesús, Daniel J. Benjamin, James O. Berger, and Thomas M. Sellke. 2016. "Rejection odds and rejection ratios: A proposal for statistical practice in testing hypotheses." *Journal of Mathematical Psychology*, 72: 90–103.

Beauchamp, Jonathan, David Cesarini, Magnus Johannesson, Matthijs J. H. M. van der Loos, Philipp D. Koellinger, et al. 2011. "Molecular Genetics And Economics." *Journal of Economic Perspectives*, 25(4): 57–82.

Beauchamp, Jonathan, Lauren Schmitz, Matt McGue, and James Lee. 2023. "Nature-Nurture Interplay: Evidence from Molecular Genetic and Pedigree Data in Korean American Adoptees." *SSRN Working Paper*. Available at: https://ssrn.com/abstract=4491976.

Becker, Gary S., and Nigel Tomes. 1976. "Child Endowments and the Quantity and Quality of Children." *Journal of Political Economy*, 84(S4): S143–S162.

Becker, Joel, Casper A. P. Burik, Grant Goldman, Nancy Wang, Hariharan Jayashankar, et al. 2021. "Resource Profile and User Guide of the Polygenic Index Repository." *Nature Human Behaviour*, 5(12): 1744–1758.

Behrman, Jere R., Robert A. Pollak, and Paul Taubman. 1982. "Parental Preferences And Provision For Progeny." *Journal of Political Economy*, 90(1): 52–73.

Belsky, Daniel W., Terrie E. Moffitt, David L. Corcoran, Benjamin Domingue, HonaLee Harrington, et al. 2016. "The Genetics of Success: How Single-Nucleotide Polymorphisms Associated With Educational Attainment Relate to Life-Course Development." *Psychological Science*, 27(7): 957–972.

Benjamin, Daniel J., Christopher F. Chabris, Edward L. Glaeser, Vilmundur Gudnason, Tamara B. Harris, David I. Laibson, Lenore J. Launder, and Shaun Purcell. 2007. "Genoeconomics." In *Biosocial Surveys.* , ed. Maxine Weinstein, James W Vaupel and Kenneth W Wachter, Chapter 15, 304–335. National Academies Press. Available from: https://www.ncbi.nlm.nih.gov/books/NBK62422/.

Benjamin, Daniel J., David Cesarini, Christopher F. Chabris, Edward L. Glaeser, David I. Laibson, et al. 2012. "The Promises and Pitfalls of Genoeconomics." *Annual Review of Economics*, 4(1): 627–662.

Bennett, Declan, Donal O'Shea, John Ferguson, Derek Morris, and Cathal Seoighe. 2021. "Controlling For Background Genetic Effects Using Polygenic Scores Improves The Power Of Genome-Wide Association Studies." *Scientific Reports*, 11(19571).

Berg, Jeremy J., Arbel Harpak, Nasa Sinnott-Armstrong, Anja Moltke Joergensen, Hakhamanesh Mostafavi, et al. 2019. "Reduced Signal for Polygenic Adaptation of Height in UK Biobank." *eLife*, 8: e39725.

Bergstrom, Theodore C. 2013. "Measures of Assortativity." *Biological Theory*, 8(2): 133–141.

Biroli, Pietro, Titus Galama, Stephanie von Hinke, Hans van Kippersluis, Cornelius A. Rietveld, and Kevin Thom. 2025. "The Economics and Econometrics of Gene–Environment Interplay." *Review of Economic Studies, in press.*, 93(1): 144–180.

Bloemendal, Alex. 2019. "A Primer on Random Matrix Theory." *YouTube Video*, Available at: https://www.youtube.com/watch?v=B7ub92OLw1g. Accessed: November 14, 2023.

Border, Richard, Athanasiadis Georgios, Buil Alfonso, Andrew J. Schork, Na Cai, et al. 2022*a*. "Cross Trait Assortative Mating Is Widespread And Inflates Genetic Correlation Estimates." *Science*, 378(6621): 754–761.

Border, Richard, Sean O'Rourke, Teresa de Candia, Michael E. Goddard, Peter M. Visscher, et al. 2022*b*. "Assortative Mating Biases Marker-based Heritability Estimators." *Nature Communications*, 13(1): 660.

Bouchard, Thomas J. 2013. "The Wilson Effect: the Increase in Heritability of IQ with Age." *Twin Research and Human Genetics*, 16(5): 923–930.

Bowden, Jack, George Davey Smith, and Stephen Burgess. 2015. "Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression." *International Journal of Epidemiology*, 44(2): 512–525.

Branigan, Amelia R., Kenneth J. McCallum, and Jeremy Freese. 2013. "Variation in the Heritability of Educational Attainment: An International Meta-Analysis." *Social Forces*, 92(1): 109–140.

Braudt, David B. 2018. "Sociogenomics in the 21st century: An introduction to the history and potential of genetically informed social science." *Sociology Compass*, 12(10).

Brumpton, Ben, Eleanor Sanderson, Karl Heilbron, Fernando Pires Hartwig, Sean Harrison, et al. 2020. "Avoiding Dynastic, Assortative Mating, And Population Stratification Biases In Mendelian Randomization Through Within-family Analyses." *Nature Communications*, 11(1): 3519.

Bulik-Sullivan, Brendan, Hilary K. Finucane, Verneri Anttila, Alexander Gusev, Felix R. Day, et al. 2015*a*. "An Atlas of Genetic Correlations across Human Diseases and Traits." *Nature Genetics*, 47(11): 1236–1241.

Bulik-Sullivan, Brendan K., Po-Ru Loh, Hilary K. Finucane, Stephan Ripke, Jian Yang, Nick Patterson, Mark J. Daly, Alkes L. Price, and Benjamin M. Neale. 2015*b*. "LD Score regression distinguishes confounding from polygenicity in genome-wide association studies." *Nature Genetics*, 47(3): 291–295.

Burgess, Stephen, Christopher N. Foley, Elias Allara, James R. Staley, and Joanna M. M. Howson. 2020*a*. "A Robust and Efficient Method for Mendelian Randomization with Hundreds of Genetic Variants." *Nature Communications*, 11(376).

Burgess, Stephen, George Davey Smith, Neil M. Davies, Frank Dudbridge, Dipender Gill, et al. 2020*b*. "Guidelines for performing Mendelian randomization investigations." *Wellcome Open Research*, 4: 186.

Burgess, Stephen, George Davey Smith, Neil M. Davies, Frank Dudbridge, Dipender Gill, et al. 2023. "Guidelines for performing Mendelian randomization investigations: update for summer 2023." *Wellcome Open Research*, 4: 186.

Bycroft, Clare, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T. Elliott, et al. 2018. "The UK Biobank resource with deep phenotyping and genomic data." *Nature*, 562(7726): 203–209.

Campos, Adrian I., Shinichi Namba, Shu-Chin Lin, Kisung Nam, Julia Sidorenko, et al. 2023. "Boosting The Power Of Genome-Wide Association Studies Within And Across Ancestries By Using Polygenic Scores." *Nature Genetics*, 55: 1769–1776.

Carvalho, Leandro S. 2025. "Genetics and Socioeconomic Status: Some Preliminary Evidence on Mechanisms." *Journal of Political Economy Microeconomics*, 3(3): 429–476.

Cawley, John, Euna Han, and Edward C. Norton. 2011. "The validity of genes related to neurotransmitters as instrumental variables." *Health Economics*, 20(8): 884–888.

Cawley, John, Euna Han, Jiyoon Kim, and Edward C. Norton. 2019. "Testing for family influences on obesity: The role of genetic nurture." *Health Economics*, 28(7): 937–952.

Cawley, John, Euna Han, Jiyoon Kim, and Edward C. Norton. 2023. "Genetic nurture in educational attainment." *Economics & Human Biology*, 49: 101239.

Cesarini, David, and Peter M. Visscher. 2017. "Genetics and educational attainment." *npj Science of Learning*, 2(1).

Chabris, Christopher F., James J. Lee, David Cesarini, Daniel J. Benjamin, and David I. Laibson. 2015. "The Fourth Law of Behavior Genetics." *Current Directions in Psychological Science*, 24(4): 304–312.

Chen, Chia-Yen, Ruoyu Tian, Tian Ge, Max Lam, Gabriela Sanchez-Andrade, et al. 2023. "The Impact Of Rare Protein Coding Genetic Variation On Adult Cognitive Function." *Nature Genetics*, 55: 927–938.

Clark, Damon, and Heather Royer. 2013. "The Effect Of Education On Adult Mortality And Health: Evidence From Britain." *American Economic Review*, 103(6): 2087–2120.

Cloninger, C. Robert, John P. Rice, and Theodor Reich. 1979. "Multifactorial Inheritance With Cultural Transmission And Assortative Mating. II. A General Model Of Combined Polygenic And Cultural Inheritance." *American Journal of Human Genetics*, 31(2): 176–198.

Conley, Dalton. 2009. "The Promise and Challenges of Incorporating Genetic Data into Longitudinal Social Science Surveys and Research." *Biodemography and Social Biology*, 55(2): 238–251.

Conley, Dalton. 2016. "Socio-Genomic Research Using Genome-Wide Molecular Data." *Annual Review of Sociology*, 42(1): 275–299.

Conley, Dalton, Thomas Laidley, Daniel W. Belsky, Jason M. Fletcher, Jason D. Boardman, and Benjamin W. Domingue. 2016. "Assortative mating and differential fertility by phenotype and genotype across the 20th century." *Proceedings of the National Academy of Sciences*, 113(24): 6647–6652.

Crow, James F., and Motoo Kimura. 1970. *An Introduction to Population Genetics Theory.* Harper & Row. Reprinted by Blackburn Press, 2009.

Cunha, Flavio, and James Heckman. 2007. "The Technology Of Skill Formation." *American Economic Review*, 97(2): 31–47.

Cunha, Flavio, and James J Heckman. 2008. "Formulating, Identifying And Estimating The Technology Of Cognitive And Noncognitive Skill Formation." *Journal of Human Resources*, 43(4): 738–782.

Daetwyler, Hans D., Beatriz Villanueva, and John A. Woolliams. 2008. "Accuracy Of Predicting The Genetic Risk Of Disease Using A Genome-wide Approach." *PLoS ONE*, 3(10): e3395.

Davey Smith, George, and Gibran Hemani. 2014. "Mendelian randomization: genetic anchors for causal inference in epidemiological studies." *Human Molecular Genetics*, 23(R1): R89–R98.

Davey Smith, George, and Shah Ebrahim. 2003. "Mendelian Randomization: Can Genetic Epidemiology Contribute to Understanding Environmental Determinants of Disease?" *International Journal of Epidemiology*, 32(1): 1–22.

Davies, Neil M., Matt Dickson, George Davey Smith, Gerard J. van den Berg, and Frank Windmeijer. 2018. "The Causal Effects of Education on Health Outcomes in the UK Biobank." *Nature Human Behaviour*, 2(2): 117–125.

de Vlaming, Ronald, Aysu Okbay, Cornelius A. Rietveld, Magnus Johannesson, Patrik K. E. Magnusson, et al. 2017. "Meta-GWAS Accuracy and Power (MetaGAP) Calculator Shows that Hiding Heritability Is Partially Due to Imperfect Genetic Correlations across Studies." *PLOS Genetics*, 13(1): e1006495.

Dias Pereira, Rita, Pietro Biroli, Titus Galama, Stephanie von Hinke, Hans van Kippersluis, Cornelius A. Rietveld, and Kevin Thom. 2022. "Gene–Environment Interplay in the Social Sciences." *Research Encyclopedia of Economics and Finance* Retrieved from: https://oxfordre.com/economics/view/10.1093/acrefore/9780190625979.001.0001/acrefore-9780190625979-e-804.

Ding, Weili, Steven F. Lehrer, J. Niels Rosenquist, and Janet Audrain-McGovern. 2009. "The Impact Of Poor Health On Academic Performance: New Evidence Using Genetic Markers." *Journal of Health Economics*, 28(3): 578–597.

Ding, Yi, Kangcheng Hou, Ziqi Xu, Aditya Pimplaskar, Ella Petter, et al. 2023. "Polygenic Scoring Accuracy Varies Across The Genetic Ancestry Continuum." *Nature*, 618: 774–781.

DiPrete, Thomas A, Casper A P Burik, and Philipp D Koellinger. 2018. "Genetic Instrumental Variable Regression: Explaining Socioeconomic And Health Outcomes In Nonexperimental Data." *Proceedings of the National Academy of Sciences*, 115(22): E4970–E4979.

Dixon, Padraig, George Davey Smith, Stephanie von Hinke, Neil M. Davies, and William Hollingworth. 2016. "Estimating Marginal Healthcare Costs Using Genetic Variants as Instrumental Variables: Mendelian Randomization in Economic Evaluation." *PharmacoEconomics*, 34(11): 1075–1086.

Domingue, Benjamin W., and Jason Fletcher. 2020. "Separating Measured Genetic and Environmental Effects: Evidence Linking Parental Genotype and Adopted Child Outcomes." *Behavior Genetics*, 50(5): 301–309.

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*, 115(4): 702–707.

Duncan, Laramie, Hanyang Shen, Bizu Gelaye, Joeri Meijsen, Kerry Ressler, Marcus Feldman, Roseann Peterson, and Benjamin Domingue. 2019. "Analysis of polygenic risk score usage and performance in diverse human populations." *Nature Communications*, 10. Article 3328.

Einav, Liran, Amy Finkelstein, Iuliana Pascu, and Mark R. Cullen. 2016. "How General Are Risk Preferences? Choices under Uncertainty in Different Domains." *American Economic Review*, 102(6): 2606–2638.

Fahed, Akl C., Anthony A. Philippakis, and Amit V. Khera. 2022. "The Potential Of Polygenic Scores To Improve Cost And Efficiency Of Clinical Trials." *Nature Communications*, 13(1): 2922.

Falconer, Douglas C. 1960. *Introduction to Quantitative Genetics.* Edinburgh and London: Oliver and Boyd.

Fang, Hai, Hans De Wolf, Bogdan Knezevic, Katie L. Burnham, Julie Osgood, et al. 2019. "A genetics-led approach defines the drug target landscape of 30 immune-related traits." *Nature Genetics*, 51(7): 1082–1091.

Fisher, Ronald A. 1918. "The Correlation between Relatives on the Supposition of Mendelian Inheritance." *Transactions of the Royal Society of Edinburgh*, 52(02): 399–433.

Fletcher, Jason. 2011. "The promises and pitfalls of combining genetic and economic research." *Health Economics*, 20(8): 889–892.

Fletcher, Jason M., and Steven F. Lehrer. 2009. "The Effects of Adolescent Health on Educational Outcomes: Causal Evidence Using Genetic Lotteries between Siblings." *Forum for Health Economics & Policy*, 12(2). Article 8.

Fletcher, Jason, Yuchang Wu, Tianchang Li, and Qiongshi Lu. 2024. "Interpreting polygenic score effects in sibling analysis." *PLOS ONE*, 19(2): e0282212.

Fowler, James H., and Christopher T. Dawes. 2013. "In Defense of Genopolitics." *American Political Science Review*, 107(2): 326–374.

Frayling, Timothy M., Nicholas J. Timpson, Michael N. Weedon, Eleftheria Zeggini, Rachel M. Freathy, et al. 2007. "A Common Variant in the FTO Gene Is Associated with Body Mass Index and Predisposes to Childhood and Adult Obesity." *Science*, 316(5826): 889–894.

Freese, Jeremy. 2018. "The Arrival of Social Science Genomics." *Contemporary Sociology: A Journal of Reviews*, 47(5): 524–536.

Frey, Renato, Andreas Pedroni, Rui Mata, Jörg Rieskamp, and Ralph Hertwig. 2017. "Risk preference shares the psychometric structure of major psychological traits." *Science Advances*, 3(10): e1701381.

Ge, Tian, Chia-Yen Chen, Yang Ni, Yen-Chen Anne Feng, and Jordan W Smoller. 2019. "Polygenic Prediction via Bayesian Regression and Continuous Shrinkage Priors." *Nature Communications*, 10(1): 1776.

Gillen, Ben, Erik Snowberg, and Leeat Yariv. 2019. "Experimenting with Measurement Error: Techniques with Applications to the Caltech Cohort Study." *Journal of Political Economy*, 127(4): 1826–1863.

Gillespie, John H. 2004. *Population Genetics: A Concise Guide.* . Second ed., The Johns Hopkins University Press.

Goldberger, Arthur S. 1978. "Models and Methods in the IQ Debate: Part I and II. Revised." Social Systems Research Institute Working Paper No. 7801.

Goldberger, Arthur S. 1991. *A Course in Econometrics.* Cambridge, MA :Harvard University Press.

Goldberger, Arthur S. 2005. "Structural Equation Models in Human Behavior Genetics." In *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg.* , ed. Donald W K Andrews and James H Stock. Cambridge University Press.

Goldberger, Arthur S AS. 1979. "Heritability." *Economica*, 46(184): 327–347.

Hamer, D., and L. Sirota. 2000. "Beware the Chopsticks Gene." *Molecular Psychiatry*, 5(1): 11–13.

Hanoch, Yaniv, Joseph G. Johnson, and Andreas Wilke. 2006. "Domain Specificity in Experimental Measures and Participant Recruitment: An Application to Risk-Taking Behavior." *Psychological Science*, 17(4): 300–304.

Haseman, Joseph K, and Robert C Elston. 1972. "The Investigation Of Linkage Between A Quantitative Trait And A Marker Locus." *Behavior Genetics*, 2(1): 3–19.

Hayes, Ben J., Peter M. Visscher, and Michael E. Goddard. 2009. "Increased accuracy of artificial selection by using the realized relationship matrix." *Genetics Research*, 91(1): 47–60.

Heckman, James J., and Sergio Urzúa. 2010. "Comparing IV with Structural Models: What Simple IV Can and Cannot Identify." *Journal of Econometrics*, 156(1): 27–37.

Herd, Pamela, Jeremy Freese, Kamil Sicinski, Benjamin W. Domingue, Kathleen Mullan Harris, Caiping Wei, and Robert M. Hauser. 2019. "Genes, Gender Inequality, and Educational Attainment." *American Sociological Review*, 84(6): 1069–1098.

Hill, William G., Michael E. Goddard, and Peter M. Visscher. 2008. "Data and theory point to mainly additive genetic variance for complex traits." *PLoS Genetics*, 4(2): e1000008.

Hivert, Valentin, Julia Sidorenko, Florian Rohart, Michael E Goddard, Jian Yang, et al. 2021. "Estimation Of Non-additive Genetic Variance In Human Complex Traits From A Large Sample Of Unrelated Individuals." *American Journal of Human Genetics*, 108(5): 786–798.

Houmark, Mikkel Aagaard, Victor Ronda, and Michael Rosholm. 2024. "The Nurture of Nature and the Nature of Nurture: How Genes and Investments Interact in the Formation of Skills." *American Economic Review*, 114(2): 385–425.

Howe, Laurence J., David M. Evans, Gibran Hemani, George Davey Smith, and Neil M. Davies. 2022*a*. "Evaluating indirect genetic effects of siblings using singletons." *PLOS Genetics*, 18(7): e1010247.

Howe, Laurence J., Michel G. Nivard, Tim T. Morris, Ailin F. Hansen, Humaira Rasheed, et al. 2022*b*. "Within-sibship genome-wide association analyses decrease bias in estimates of direct genetic effects." *Nature Genetics*, 54(5): 581–592.

Hwang, Liang-Dar, Justin D. Tubbs, Justin Luong, Mischa Lundberg, Gunn-Helen Moen, et al. 2020. "Estimating indirect parental genetic effects on offspring phenotypes using virtual parental genotypes derived from sibling and half sibling pairs." *PLOS Genetics*, 16(10): e1009154.

Imbens, Guido W. 2010. "Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzúa (2009)." *Journal of Economic Literature*, 48(2): 399–423.

Imbens, Guido W., and Joshua D. Angrist. 1996. "Identification and Estimation of Local Average Treatment Effects." *Econometrica*, 62(2): 467–475.

Jami, Eshim S., Anke R. Hammerschlag, Meike Bartels, and Christel M. Middeldorp. 2021. "Parental characteristics and offspring mental health and related outcomes: a systematic review of genetically informative literature." *Translational Psychiatry*, 11(1).

Jencks, Christopher. 1980. "Heredity, Environment, and Public Policy Reconsidered." *American Sociological Review*, 45(5): 723–736.

Jencks, Christopher, and Marsha Brown. 1977. "Genes and Social Stratification: A Methodological Exploration with Illustrative Data." In *Kinometrics: Determinants of Socioeconomic Success Within and Between Families.* , ed. Paul Taubman, 178–233. Amsterdam, New York, Oxford:North-Holland Publishing Company.

Johnson, Rebecca, Ramina Sotoudeh, and Dalton Conley. 2022. "Polygenic Scores for Plasticity: A New Tool for Studying Gene–Environment Interplay." *Demography*, 59(3): 1045–1070.

Jurgens, Sean J., James P. Pirruccello, Seung Hoan Choi, Valerie N. Morrill, Mark Chaffin, et al. 2023. "Adjusting For Common Variant Polygenic Scores Improves Yield In Rare Variant Association Analyses." *Nature Genetics*, 55(4): 544–548.

Karlsson Linnér, Richard, Pietro Biroli, Edward Kong, S. Fleur W. Meddens, Robbee Wedow, et al. 2019. "Genome-wide Association Analyses Of Risk Tolerance And Risky Behaviors In Over 1 Million Individuals Identify Hundreds Of Loci And Shared Genetic Influences." *Nature Genetics*, 51(2): 245–257.

Khera, Amit V., Mark Chaffin, Kaitlin H. Wade, Sohail Zahid, Joseph Brancale, et al. 2019. "Polygenic Prediction of Weight and Obesity Trajectories from Birth to Adulthood." *Cell*, 177(3): 587–596.e9.

Khera, Amit V., Mark Chaffin, Krishna G. Aragam, Mary E. Haas, Carolina Roselli, et al. 2018. "Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations." *Nature Genetics*, 50(9): 1219–1224.

Kong, Augustine, Gudmar Thorleifsson, Michael L. Frigge, Bjarni Vilhjálmsson, Alexander I. Young, et al. 2018. "The nature of nurture: Effects of parental genotypes." *Science*, 359(6374): 424–428.

Lambert, Jean-Charles, Carla A Ibrahim-Verbaas, Denise Harold, Adam C Naj, Rebecca Sims, et al. 2013. "Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease." *Nature Genetics*, 45(12): 1452–1458.

Lander, Eric S., and Nicholas J. Schork. 1994. "Genetic Dissection of Complex Traits." *Science*, 265(5181): 2037–2048.

Lee, James J., and Carson C. Chow. 2013. "The Causal Meaning of Fisher's Average Effect." *Genetics Research*, 95(2–3): 89–109.

Lee, James J., Robbee Wedow, Aysu Okbay, Edward Kong, Omeed Maghzian, et al. 2018. "Gene Discovery And Polygenic Prediction From A 1.1-million-person GWAS Of Educational Attainment." *Nature Genetics*, 50(8): 1112–1121.

Lloyd-Jones, Luke R., Jian Zeng, Julia Sidorenko, Loïc Yengo, Gerhard Moser, et al. 2019. "Improved polygenic prediction by Bayesian multiple regression on summary statistics." *Nature Communications*, 10(1): 5086.

Locke, Adam E, Bratati Kahali, Sonja I Berndt, Anne E Justice, Tune H Pers, et al. 2015. "Genetic studies of body mass index yield new insights for obesity biology." *Nature*, 518(7538): 197–206.

Loehlin, John C. 1978. "Heredity-environment analyses of Jencks's IQ correlations." *Behavior Genetics*, 8(5): 415–436.

Loehlin, John C. 2009. "History of Behavior Genetics." In *Handbook of Behavior Genetics*. , ed. Yong-Kyu Kim. Springer New York.

Maniadis, Zacharias, Fabio Tufano, and John A. List. 2014. "One Swallow Doesn't Make a Summer: New Evidence on Anchoring Effects." *American Economic Review*, 104(1): 277–290.

Manski, Charles F. 1993. "Identification of endogenous social effects: The reflection problem." *The Review of Economic Studies*, 60(3): 531–542.

Manski, Charles F. 2011. "Genes, Eyeglasses, and Social Policy." *Journal of Economic Perspectives*, 25(4): 83–94.

Markel, Gareth, Jonathan Beauchamp, Rafael Ahlskog, Joakim Ebeltoft Coleman, René Möttus, Sven Oskarsson, Uku Vainik, and Eivind Ystrøm. 2025. "Nature, nurture, and socioeconomic outcomes: New evidence from sib pairs and molecular genetic data." *SSRN Working Paper*. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5225447.

Martin, Alicia R., Christopher R. Gignoux, Raymond K. Walters, Genevieve L. Wojcik, Benjamin M. Neale, et al. 2017. "Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations." *American Journal of Human Genetics*, 100(4): 635–649.

Martin, Alicia R., Masahiro Kanai, Yoichiro Kamatani, Yukinori Okada, Benjamin M. Neale, et al. 2019. "Clinical Use Of Current Polygenic Risk Scores May Exacerbate Health Disparities." *Nature Genetics*, 51(4): 584–591.

Martin, Joanna, Ekaterina A. Khramtsova, Slavina B. Goleva, Gabriëlla A. M. Blokland, Michela Traglia, et al. 2021. "Examining Sex-Differentiated Genetic Effects Across Neuropsychiatric and Behavioral Traits." *Biological Psychiatry*, 89(12): 1127–1137.

Martschenko, Daphne Oluwaseun, Benjamin W Domingue, Lucas J Matthews, and Sam Trejo. 2021. "FoGS provides a public FAQ repository for social and behavioral genomic discoveries." *Nature Genetics*, 53(9): 1272–1274.

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).

McMartin, Andrew, and Dalton Conley. 2020. "Commentary: Mendelian Randomization and Education–Challenges Remain." *International Journal of Epidemiology*, 49(4): 1193–1206.

Mendel, Gregor. 1866. "Versuche über Pflanzen-Hybriden." *Verhandlungen des naturforschenden Vereines in Brünn*, 4: 3–47.

Menozzi, Paolo, Alberto Piazza, and Luigi Cavalli-Sforza. 1978. "Synthetic Maps of Human Gene Frequencies in Europeans." *Science*, 201(4358): 786–792.

Meyer, Michelle N, Paul S Appelbaum, Daniel J Benjamin, Shawneequa L Callier, Nathaniel Comfort, et al. 2023. "Wrestling with Social and Behavioral Genomics: Risks, Potential Benefits, and Ethical Responsibility." *Hastings Center Report*, 53: S2–S49.

Miao, Jiacheng, Gefei Song, Yixuan Wu, Jiaxin Hu, Yuchang Wu, Shubhashrita Basu, James S. Andrews, Katherine Schaumberg, Jason M. Fletcher, Lauren L. Schmitz, and Qiongshi Lu. 2022*a*. "Reimagining Gene-Environment Interaction Analysis for Human Complex Traits." *bioRχiv*. Available at: https://www.biorxiv.org/content/10.1101/2022.12.11.519973v1.

Miao, Jiacheng, Gefei Song, Yixuan Wu, Jiaxin Hu, Yuchang Wu, Shubhashrita Basu, James S. Andrews, Katherine Schaumberg, Jason M. Fletcher, Lauren L. Schmitz, and Qiongshi Lu. 2025. "PIGEON: a statistical framework for estimating gene–environment interaction for polygenic traits." *Nature Human Behaviour*, 9(8): 1654–1668.

Miao, Jiacheng, Hanmin Guo, Gefei Song, Zijie Zhao, Lin Hou, et al. 2022*b*. "Quantifying Portable Genetic Effects And Improving Cross-ancestry Genetic Prediction With GWAS Summary Statistics." *Nature Communications*, 14(832).

Mills, Melinda C, and Charles Rahal. 2019. "A Scientometric Review Of Genome-wide Association Studies." *Communications Biology*, 2(1): 9.

Mills, Melinda C., and Charles Rahal. 2020. "The GWAS Diversity Monitor Tracks Diversity By Disease In Real Time." *Nature Genetics*, 52(3): 242–243.

Mills, Melinda C., and Felix C. Tropf. 2020. "Sociology, Genetics, and the Coming of Age of Sociogenomics." *Annual Review of Sociology*, 46(1): 553–581.

Mills, Melinda C., Nicola Barban, and Felix C. Tropf. 2020. *An Introduction to Statistical Genetic Data Analysis.* The MIT Press.

Millwood, Iona Y, Robin G Walters, Xue W Mei, Yu Guo, Ling Yang, et al. 2019. "Conventional and Genetic Evidence on Alcohol and Vascular Disease Aetiology: A Prospective Study of 500,000 Men and Women in China." *The Lancet*, 393(10183): 1831–1842.

Mostafavi, Hakhamanesh, Arbel Harpak, Ipsita Agarwal, Dalton Conley, Jonathan K Pritchard, et al. 2020. "Variable prediction accuracy of polygenic scores within an ancestry group." *eLife*, 9: e48376.

Muglia, Pierandrea, Federica Tozzi, Nicholas W. Galwey, Clyde Francks, Ruchi Upmanyu, et al. 2008. "Genome-wide association study of recurrent major depressive disorder in two European case–control cohorts." *Molecular Psychiatry*, 15(6): 589–601.

Munafò, Marcus R., and George Davey Smith. 2018. "Robust research needs many lines of evidence." *Nature*, 553(7689): 399–401.

Nivard, Michel G., Daniel W. Belsky, K. Paige Harden, Tina Baier, Ole A. Andreassen, Eivind Ystrøm, Elsje van Bergen, and Torkild H.

Lyngstad. 2024. "More than nature and nurture: indirect genetic effects on children's academic achievement are consequences of dynastic social processes." *Nature Human Behaviour*, 8(4): 771–778.

Norton, Edward, and Euna Han. 2008. "Genetic Information, Obesity and Labor Market Outcomes." *Health Economics*, 17: 1089–1104.

Novembre, John, Toby Johnson, Katarzyna Bryc, Zoltán Kutalik, Adam R Boyko, et al. 2008. "Genes Mirror Geography Within Europe." *Nature*, 456(7218): 98–101.

Nurk, Sergey, Sergey Koren, Arang Rhie, Mikko Rautiainen, Andrey V. Bzikadze, et al. 2022. "The Complete Sequence Of A Human Genome." *Science*, 376(6558): 44–53.

Okbay, Aysu, Jonathan P. Beauchamp, Mark A Fontana, James J Lee, Tune H Pers, et al. 2016. "Genome-wide association study identifies 74 loci associated with educational attainment." *Nature*, 533(7604): 539–542.

Okbay, Aysu, Yeda Wu, Nancy Wang, Hariharan Jayashankar, Michael Bennett, et al. 2022. "Polygenic prediction of educational attainment within and between families from genome-wide association analyses in 3 million individuals." *Nature Genetics 2022 54:4*, 54(4): 437–449.

Panagiotou, Orestis A., and John P. A. Ioannidis. 2012. "What Should The Genome-wide Significance Threshold Be? Empirical Replication Of Borderline Genetic Associations." *International Journal of Epidemiology*, 41(1): 273–286.

Papageorge, Nicholas W, and Kevin Thom. 2020. "Genes, Education, and Labor Market Outcomes: Evidence from the Health and Retirement Study." *Journal of the European Economic Association*, 18(3): 1351–1399.

Patterson, Nick, Alkes L Price, and David Reich. 2006. "Population Structure and Eigenanalysis." *PLoS Genetics*, 2(12): e190.

Pazokitoroudi, Ali, Alec M. Chiu, Kathryn S. Burch, Bogdan Pasaniuc, and Sriram Sankararaman. 2021. "Quantifying the contribution of dominance deviation effects to complex trait variation in biobank-scale data." *American Journal of Human Genetics*, 108(5): 799–808.

Plomin, Robert, John C. DeFries, and John C. Loehlin. 1977. "Genotype-environment Interaction And Correlation In The Analysis Of Human Behavior." *Psychological Bulletin*, 84(2): 309–322.

Plomin, Robert, John C. DeFries, Valerie S. Knopik, and Jenae M. Neiderhiser. 2016. "Top 10 Replicated Findings From Behavioral Genetics." *Perspectives in Psychological Science*, 11(1): 3–23.

Price, Alkes L., Nick J. Patterson, Robert M. Plenge, Michael E. Weinblatt, Nancy A. Shadick, et al. 2006. "Principal Components Analysis Corrects For Stratification In Genome-Wide Association Studies." *Nature Genetics*, 38(8): 904–909.

Purcell, Shaun M., Naomi R. Wray, Jennifer L. Stone, Peter M. Visscher, Michael C. O'Donovan, et al. 2009. "Common Polygenic Variation Contributes To Risk Of Schizophrenia And Bipolar Disorder." *Nature*, 460(7256): 748–752.

Rietveld, Cornelius A, Sarah E Medland, Jaime Derringer, Jian Yang, and Tõnu Esko et al. 2013. "GWAS of 126,559 individuals identifies genetic variants associated with educational attainment." *Science*, 340(6139): 1467–1471.

Rimfeld, Kaili, Eva Krapohl, Maciej Trzaskowski, Jonathan R. I. Coleman, Saskia Selzam, et al. 2018. "Genetic Influence On Social Outcomes During And After The Soviet Era In Estonia." *Nature Human Behaviour*, 2(4): 269–275.

Robinson, Gene E., Christina M. Grozinger, and Charles W. Whitfield. 2005. "Sociogenomics: social life in molecular terms." *Nature Reviews Genetics*, 6: 257–270.

Robinson, Matthew R., Aaron Kleinman, Mariaelisa Graff, Anna A. E. Vinkhuyzen, David Couper, et al. 2017. "Genetic evidence of assortative mating in humans." *Nature Human Behaviour*, 1. Article number: 0016.

Roura-Monllor, Jaime A., Zachary Walker, Joel M. Reynolds, Greysha Rivera-Cruz, Avner Hershlag, et al. 2025. "Promises and pitfalls of preimplantation genetic testing for polygenic disorders: a narrative review." *F&S Review*, 6(1): 100085.

Ruan, Yunfeng, Yen-Feng Lin, Yen-Chen Anne Feng, Chia-Yen Chen, Max Lam, et al. 2022. "Improving Polygenic Prediction In Ancestrally Diverse Populations." *Nature Genetics*, 54(5): 573–580.

Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology*, 66(5): 688.

Rustichini, Aldo, William G. Iacono, James J. Lee, and Matt McGue. 2023. "Educational Attainment and Intergenerational Mobility: A Polygenic Score Analysis." *Journal of Political Economy*, 131(10): 2724–2779.

Rutherford, Adam. 2022. *Control: The Dark History and Troubling Present of Eugenics.* WW Norton & Company.

Sacerdote, Bruce. 2011. "Nature and Nurture Effects On Children's Outcomes: What Have We Learned From Studies of Twins And Adoptees?" In *Handbook of Social Economics.* , ed. Jess Benhabib, Alberto. Bisin and Matthew O. Jackson, Chapter 1, 1–29. Elsevier/North-Holland.

Sanz-de Galdeano, Anna, and Anastasia Terskaya. 2025. "Sibling Differences in Genetic Propensity for Education: How do Parents React?" *The Review of Economics and Statistics*, 107(5): 1356–1370.

Schmitz, Lauren L., and Dalton Conley. 2017. "The effect of Vietnam-era conscription and genetic potential for educational attainment on schooling outcomes." *Economics of Education Review*, 61: 85–97.

Schmitz, Lauren L., Julia Goodwin, Jiacheng Miao, Qiongshi Lu, and Dalton Conley. 2021. "The Impact Of Late-Career Job Loss And Genetic Risk On Body Mass Index: Evidence From Variance Polygenic Scores." *Scientific Reports*, 11(7647).

Schunkert, Heribert, Emanuele Di Angelantonio, Michael Inouye, Riyaz S Patel, Samuli Ripatti, et al. 2025. "Clinical utility and implementation of polygenic risk scores for predicting cardiovascular disease." *European Heart Journal*, 46(15): 1372–1383.

Shen, Hao, and Marcus W. Feldman. 2020. "Genetic nurturing, missing heritability, and causal analysis in genetic statistics." *Proceedings of the National Academy of Sciences*, 117(41): 25646–25654.

Silventoinen, Karri, Sampo Sammalisto, Markus Perola, Dorret I. Boomsma, Belinda K. Cornes, et al. 2003. "Heritability of adult body height: a comparative study of twin cohorts in eight countries." *Twin Research and Human Genetics*, 6(5): 399–408.

Sohail, Mashaal, Robert M. Maier, Andrea Ganna, Alex Bloemendal, Alicia R. Martin, et al. 2019. "Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies." *eLife*, 8: e39702.

Sotoudeh, Ramina, Kathleen Mullan Harris, and Dalton Conley. 2019. "Effects of the peer metagenomic environment on smoking behavior." *Proceedings of the National Academy of Sciences*, 116(33): 16302–16307.

Speed, Doug, and David J. Balding. 2019. "SumHer better estimates the SNP heritability of complex traits from summary statistics." *Nature Genetics*, 51: 277–284.

Stefansson, Hreinn, Roel A. Ophoff, Stacy Steinberg, Ole A. Andreassen, Sven Cichon, et al. 2009. "Common variants conferring risk of schizophrenia." *Nature*, 460(7256): 744–747.

Strachan, Tom, and Andrew P Read. 2018. *Human Molecular Genetics.* . 5th ed., Garland Science.

Sved, John A, and William G Hill. 2018. "One hundred years of linkage disequilibrium." *Genetics*, 209(3): 629–636.

Tahmasbi, Rasool, Luke M. Evans, Eric Turkheimer, and Matthew C. Keller. 2017. "Testing the moderation of quantitative gene by environment interactions in unrelated individuals." *bioRχiv*. Available at: https://www.biorxiv.org/content/10.1101/191080v1.

Tan, Tammy, Hariharan Jayashankar, Junming Guan, S. Moeen Nehzati, et al. 2024. "Family-GWAS Reveals Effects of Environment and Mating

on Genetic Associations." *medRxiv*. Available at: https://www.medrxiv.org/content/10.1101/2024.10.01.24314703v1.

Taubman, Paul. 1976. "The Determinants Of Earnings: Genetics, Family, And Other Environments: A Study Of White Male Twins." *American Economic Review*, 66(5): 858–870.

Trejo, Sam, and Benjamin W Domingue. 2018. "Genetic nature or genetic nurture? Introducing social genetic parameters to quantify bias in polygenic score analyses." *Biodemography and Social Biology*, 64(3-4): 187–215.

Trejo, Sam, and Klint Kanopka. 2024. "Using the phenotype differences model to identify genetic effects in samples of partially genotyped sibling pairs." *Proceedings of the National Academy of Sciences*, 121(49).

Tucker-Drob, Elliot M. 2017. "Measurement Error Correction of Genome-Wide Polygenic Scores in Prediction Samples." *bioRχiv*. Available at: https://www.biorxiv.org/content/10.1101/165472v1.

Turley, Patrick, Alicia R. Martin, Grant Goldman, Hui Li, Masahiro Kanai, et al. 2021. "Multi-Ancestry Meta-Analysis Yields Novel Genetic Discoveries And Ancestry-Specific Associations." *bioRχiv*. Available at: https://www.biorxiv.org/content/10.1101/2021.04.23.441003v1.

van Kippersluis, Hans, Pietro Biroli, Rita Dias Pereira, Titus J. Galama, Stephanie von Hinke, et al. 2023. "Overcoming attenuation bias in regressions using polygenic indices." *Nature Communications*, 14(4473). Article number: 4473.

Veller, Carl, and Graham M Coop. 2024. "Interpreting population- and family-based genome-wide association studies in the presence of confounding." *PLoS Biology*, 22(4): e3002511.

Veller, Carl, Molly Przeworski, and Graham Coop. 2024. "Causal Interpretations Of Family GWAS In The Presence Of Heterogeneous Effects." *Proceedings of the National Academy of Sciences of the United States of America*, 121(38): e2401379121.

Vilhjálmsson, Bjarni J., Jian Yang, Hilary K. Finucane, Alexander Gusev, Sara Lindström, et al. 2015. "Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores." *The American Journal of Human Genetics*, 97(4): 576–592.

Visscher, Peter M. 2008. "Sizing up human height variation." *Nature Genetics*, 40(5): 489–490.

Visscher, Peter M., Matthew A. Brown, Mark I. McCarthy, and Jian Yang. 2012. "Five years of GWAS Discovery." *The American Journal of Human Genetics*, 90(1): 7–24.

Visscher, Peter M., Naomi R. Wray, Qian Zhang, Pamela Sklar, Mark I. McCarthy, et al. 2017. "10 Years of GWAS Discovery: Biology, Function, and Translation." *The American Journal of Human Genetics*, 101(1): 5–22.

Visscher, Peter M, Sarah E Medland, Manuel A R Ferreira, Katherine I Morley, Gu Zhu, et al. 2006. "Assumption-free Estimation Of Heritability From Genome-wide Identity-by-descent Sharing Between Full Siblings." *PLoS Genetics*, 2(3): e41.

Visscher, Peter, William Hill, and Naomi Wray. 2008. "Heritability in the Genomics Era - Concepts and Misconceptions." *Nature Reviews Genetics*, 9: 255–266.

von Hinke Kessler Scholder, Stephanie, George Davey Smith, Debbie A Lawlor, Carol Propper, and Frank Windmeijer. 2011. "Mendelian randomization: the use of genes in instrumental variable analyses." *Health Economics*, 20(8): 893–896.

von Hinke Kessler Scholder, Stephanie, George Davey Smith, Debbie A. Lawlor, Carol Propper, and Frank Windmeijer. 2013. "Child height, health and human capital: Evidence using genetic markers." *European Economic Review*, 57: 1–22.

von Hinke, Stephanie, George Davey Smith, Debbie A. Lawlor, Carol Propper, and Frank Windmeijer. 2016. "Genetic Markers as Instrumental Variables." *Journal of Health Economics*, 45: 131–148.

Walsh, Bruce, and Michael Lynch. 2018. "Associative Effects: Competition, Social Interactions, Group and Kin Selection." In *Evolution and Selection of Quantitative Traits*. Chapter 22. Oxford:Oxford University Press.

Wang, Huanwei, Futao Zhang, Jian Zeng, Yang Wu, Kathryn E. Kemper, et al. 2019. "Genotype-by-environment interactions inferred from genetic effects on phenotypic variability in the UK Biobank." *Science Advances*, 5(8).

Wang, Xin, Sotiris Karagounis, Suyash S. Shringarpure, Rohith Srivas, Qiaojuan Jane Su, Vladimir Vacic, Steven J. Pitts, and Adam Auton. 2024. "The Impact on Clinical Success from the 23andMe Cohort." *medRxiv*. Available at: https://www.medrxiv.org/content/10.1101/2024.06.17.24309059v1.

Wang, Ying, Jing Guo, Guiyan Ni, Jian Yang, Peter M. Visscher, et al. 2020. "Theoretical and empirical quantification of the accuracy of polygenic scores in ancestry divergent populations." *Nature Communications*, 11(1): 1–9.

Weber, Elke U, Ann-rené E Blais, and Nancy E Betz. 2002. "A Domain-Specific Risk-Attitude Scale: Measuring Risk Perceptions and Risk Behaviors." *Journal of Behavioral Decision Making*, 15(4): 263–290.

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*, 83(4): 802–832.

Wellcome Trust Case Control Consortium. 2007. "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls." *Nature*, 447(7145): 661–678.

Wetterstrand, KA. 2023. "DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP)." *Available at: www.genome.gov/sequencingcostsdata. Accessed April 4, 2023.*

Widen, Erik, Timothy G. Raben, Louis Lello, and Stephen D. H. Hsu. 2021. "Machine Learning Prediction of Biomarkers from SNPs and of Disease Risk from Biomarkers in the UK Biobank." *Genes*, 12(7).

Wientjes, Yvonne C. J., Piter Bijma, Roel F. Veerkamp, and Mario P. L. Calus. 2016. "An Equation to Predict the Accuracy of Genomic Values by Combining Data from Multiple Traits, Populations, or Environments." *Genetics*, 202(2): 799–823.

Wientjes, Yvonne C. J., Roel F. Veerkamp, Piter Bijma, Henk Bovenhuis, Chris Schrooten, and Mario P. L. Calus. 2015. "Empirical And Deterministic Accuracies Of Across-Population Genomic Prediction." *Genetics Selection Evolution*, 47(5): 1–14.

Winkler, Thomas W., Felix R. Day, Damien C. Croteau-Chonka, Andrew R. Wood, Adam E. Locke, et al. 2014. "Quality Control And Conduct Of Genome-Wide Association Meta-Analyses." *Nature Protocols*, 9(5): 1192–1212.

Wray, Naomi R., Jian Yang, Ben J. Hayes, Alkes L. Price, Michael E. Goddard, et al. 2013. "Pitfalls of predicting complex traits from SNPs." *Nature Reviews Genetics*, 14(7): 507–15.

Wray, Naomi R., Michael E. Goddard, and Peter M. Visscher. 2007. "Prediction of individual genetic risk to disease from genome-wide association studies." *Genome Research*, 17(10): 1520–1528.

Wu, Yang, Zhili Zheng, Peter M. Visscher, and Jian Yang. 2017. "Quantifying the mapping precision of genome-wide association studies using whole-genome sequencing data." *Genome Biology*, 18(1): 86.

Yang, Jian, Beben Benyamin, Brian P. McEvoy, Scott Gordon, Anjali K. Henders, et al. 2010. "Common SNPs explain a large proportion of the heritability for human height." *Nature Genetics*, 42(7): 565–569.

Yengo, Loïc, Matthew R. Robinson, Matthew C. Keller, Kathryn E. Kemper, Yuanhao Yang, et al. 2018. "Imprint of Assortative Mating on the Human Genome." *Nature Human Behaviour*, 2(12): 948–954.

Yengo, Loïc, Sailaja Vedantam, Eirini Marouli, Julia Sidorenko, Eric Bartell, et al. 2022. "A Saturated Map Of Common Genetic Variants Associated With Human Height." *Nature*, 610: 704–712.

Young, Alexander I., Michael L. Frigge, Daniel F. Gudbjartsson, Gudmar Thorleifsson, Gyda Bjornsdottir, et al. 2018. "Relatedness Dise-

quilibrium Regression Estimates Heritability Without Environmental Bias." *Nature Genetics*, 50(9): 1304–1310.

Young, Alexander I., Seyed Moeen Nehzati, Stefania Benonisdottir, Aysu Okbay, Hariharan Jayashankar, et al. 2022. "Mendelian Imputation Of Parental Genotypes Improves Estimates Of Direct Genetic Effect." *Nature Genetics*, 54(6): 897–905.

Young, Alexander Strudwick. 2023. "Estimation Of Indirect Genetic Effects And Heritability Under Assortative Mating." bioRχiv. Available at https://www.biorxiv.org/content/10.1101/2023.07.10.548458v1.

Zaidi, Arslan A., and Iain Mathieson. 2020. "Demographic history mediates the effect of stratification on polygenic scores." *eLife*, 9: e61548.

Zhang, Qianqian, Florian Privé, Bjarni Vilhjálmsson, and Doug Speed. 2021. "Improved Genetic Prediction of Complex Traits from Individual-Level Data or Summary Statistics." *Nature Communications*, 12(1): 1–9.

Zhao, Zijie, Jie Song, Tuo Wang, and Qiongshi Lu. 2021. "Polygenic Risk Scores: Effect Estimation and Model Optimization." *Quantitative Biology*, 9(2): 133–140.

Zhao, Zijie, Xiaoyu Yang, Stephen Dorn, Jiacheng Miao, Silvia H. Barcellos, Jason M. Fletcher, and Qiongshi Lu. 2025. "Controlling for polygenic genetic confounding in epidemiologic association studies." *Proceedings of the National Academy of Sciences*, 121(44): e2408715121.

"Technical Appendix"

January 2026

I Local Average Treatment Effect (LATE)

Here, we show that the two-stage least squares (2SLS) estimate of the effect of a SNP on a phenotype, instrumenting the child's genotype with the deviation from the mean parental genotype, produces a local average treatment effect (LATE) with weights equal to the child's number of heterozygous parents for that SNP. It is the same LATE as that of the effect of a SNP from a trio design and—assuming there are no **interpersonal** genetic effects from siblings—that of a sibling-difference design. **For a sibling-difference design and under the assumption of no interpersonal genetic effects from siblings, Veller, Przeworski and Coop (2024) previously derived the expressions below, and many of the steps below in our 2SLS derivation are nearly identical to those in Veller, Przeworski and Coop (2024).**

Consider a setting where SNP effects are heterogeneous, such that

$$y_i = x_i\beta_i + x_{p,i}\gamma_i + \xi_i,$$

where y_i is the phenotype, x_i is the genotype, β_i is the causal effect of that SNP for person i, $x_{p,i}$ is the **mean** parental genotype, γ_i is the coefficient on **mean** parental genotype for person i, and ξ_i is the residual. For concreteness, we assume that the SNP is conditionally uncorrelated with other genetic variants, though we could define β_i as the GWAS coefficient β_i^{GWAS} and all subsequent results would hold. Due to Mendelian segregation, we can decompose a person's genotype into

$$x_i = x_{p,i} + x_{r,i},$$

where $x_{r,i}$ is the random component of person i's genotype. The coefficient of a just-identified 2SLS estimate is

$$\beta_{2sls} = \frac{\mathrm{Cov}\left(y_i, x_{r,i}\right)/\mathrm{Var}\left(x_{r,i}\right)}{\mathrm{Cov}\left(x_i, x_{r,i}\right)/\mathrm{Var}\left(x_{r,i}\right)} = \frac{\mathrm{Cov}\left(y_i, x_{r,i}\right)}{\mathrm{Cov}\left(x_i, x_{r,i}\right)},$$

where the denominator corresponds to the "first stage" and the numerator corresponds to the "reduced form." Beginning with the first-stage term of this ratio, we evaluate

$$\begin{aligned}\mathrm{Cov}\left(x_i, x_{r,i}\right) &= \mathrm{Cov}\left(x_{p,i} + x_{r,i}, x_{r,i}\right)\\ &= \mathrm{Cov}\left(x_{p,i}, x_{r,i}\right) + \mathrm{Var}\left(x_{r,i}\right)\\ &= \mathrm{Var}\left(x_{r,i}\right) = \int x_{r,i}^2 dF_{x_{r,i}},\end{aligned}$$

where $F_{x_{r,i}}$ is the population distribution of $x_{r,i}$.

Next, we evaluate the reduced-form term and obtain

$$\begin{aligned}\text{Cov}\left(y_i, x_{r,i}\right) &= \text{Cov}\left(x_i\beta_i + x_{p,i}\gamma_i + \xi_i, x_{r,i}\right) \\ &= \text{Cov}\left(x_{r,i}\beta_i + x_{p,i}\left(\beta_i + \gamma_i\right) + \xi_i, x_{r,i}\right) \\ &= \text{Cov}\left(x_{r,i}\beta_i, x_{r,i}\right) + \text{Cov}\left(x_{p,i}\left(\beta_i + \gamma_i\right), x_{r,i}\right) + \text{Cov}\left(\xi_i, x_{r,i}\right) \\ &= \text{Cov}\left(x_{r,i}\beta_i, x_{r,i}\right) = \int \beta_i x_{r,i}^2 dF_{x_{r,i}}.\end{aligned}$$

Combining these expressions gives us

$$\beta_{2sls} = \frac{\int \beta_i x_{r,i}^2 dF_{x_{r,i}}}{\int x_{r,i}^2 dF_{x_{r,i}}}.$$

Thus, the coefficient on the child's genotype in a regression of the phenotype onto both the child's and parental genotypes yields a weighted average of the causal effects of the genotypes, weighted by the squared random component of a child's genotype. We can build additional intuition for this expression by splitting the sample into three sets, H_0, H_1, and H_2, corresponding to the individuals who have zero, one, or two heterozygous parents. Notice that if an individual has no heterozygous parents, then they will always have a genotype equal to the mean parental genotype, so $x_{r,i}^2 = 0$ for all i. For individuals with one heterozygous parent, $x_{r,i} \in \left\{-\frac{1}{2}, \frac{1}{2}\right\}$ and therefore $x_{r,i}^2 = \frac{1}{4}$ for all i. For individuals with two heterozygous parents, $x_{r,i} \in \{-1, 0, 1\}$ with probabilities of $\frac{1}{4}$, $\frac{1}{2}$, and $\frac{1}{4}$ for each element, respectively. This means that $x_{r,i}^2 \in \{0, 1\}$ with a probability of $\frac{1}{2}$ for

each state and hence $\mathbb{E}\left(x_{r,i}^2|H_2\right) = \frac{1}{2}$. Thus,

$$\begin{aligned}
\beta_{2sls} &= \frac{\int \beta_i x_{r,i}^2 dF_{x_{r,i}}}{\int x_{r,i}^2 dF_{x_{r,i}}} = \frac{\mathbb{E}_{x_{r,i}}\left(\beta_i x_{r,i}^2\right)}{\mathbb{E}_{x_{r,i}}\left(x_{r,i}^2\right)} \\
&= \frac{\mathbb{E}_{x_{r,i}}\left(\beta_i x_{r,i}^2|H_0\right)\pi_0 + \mathbb{E}_{x_{r,i}}\left(\beta_i x_{r,i}^2|H_1\right)\pi_1 + \mathbb{E}_{x_{r,i}}\left(\beta_i x_{r,i}^2|H_2\right)\pi_2}{\mathbb{E}_{x_{r,i}}\left(x_{r,i}^2\right)} \\
&= \frac{\mathbb{E}_{x_{r,i}}\left(\beta_i|H_1\right)\mathbb{E}_{x_{r,i}}\left(x_{r,i}^2|H_1\right)\pi_1 + \mathbb{E}_{x_{r,i}}\left(\beta_i|H_2\right)\mathbb{E}_{x_{r,i}}\left(x_{r,i}^2|H_2\right)\pi_2}{\mathbb{E}_{x_{r,i}}\left(x_{r,i}^2\right)} \\
&= \frac{\frac{1}{4}\pi_1\mathbb{E}_{x_{r,i}}\left(\beta_i|H_1\right) + \frac{1}{2}\pi_2\mathbb{E}_{x_{r,i}}\left(\beta_i|H_2\right)}{\mathbb{E}_{x_{r,i}}\left(x_{r,i}^2\right)} \\
&= \frac{\frac{1}{4}\pi_1\mathbb{E}_{x_{r,i}}\left(\beta_i|H_1\right) + \frac{1}{2}\pi_2\mathbb{E}_{x_{r,i}}\left(\beta_i|H_2\right)}{\frac{1}{4}\pi_1 + \frac{1}{2}\pi_2} \\
&= \frac{\pi_1}{\pi_1 + 2\pi_2}\mathbb{E}_{x_{r,i}}\left(\beta_i|H_1\right) + \frac{2\pi_2}{\pi_1 + 2\pi_2}\mathbb{E}_{x_{r,i}}\left(\beta_i|H_2\right),
\end{aligned}$$

where π_0, π_1, and π_2 are the fraction of individuals with zero, one, and two heterozygous parents, respectively.

This expression makes clear a few key points. First, individuals with homozygous parents receive no weight in this regression. So to the degree that individuals with homozygous parents have systematically different genetic effect sizes, family-based estimates will not generalize to those individuals. Second, individuals with two heterozygous parents receive double the weight as those with one heterozygous parent. Third, in genetic studies with diverse-ancestry samples, particular ancestry groups will get more weight for some genetic variants than others. That is because certain genotypes will be more common in certain groups. **As a result, even if the sample is relatively balanced between the different ancestry groups, an ancestry group with genotype frequencies closer to one-half will tend to have relatively more individuals with one or two heterozygous parents, so the estimated average effect for that genetic variant will give more weight to that ancestry group.**

II Derivations of Formulae for PGI Predictive Power

Here, we derive analytic formulae for the predictive power of a PGI, beginning with Equation (7) in the main text (de Vlaming et al., 2017). Letting $\breve{g}_{pred}$ denote the optimal predictor in the prediction population:

$$
\begin{aligned}
R^2 &= \frac{\mathrm{Cov}\big(y_{\mathrm{pred}},\hat{g}\big)^2}{\mathrm{Var}\big(y_{\mathrm{pred}}\big)\ \mathrm{Var}\big(\hat{g}\big)} = \frac{\mathrm{Cov}\Big(y_{\mathrm{pred}},\dfrac{\breve{g}+e}{\mathrm{sd}(\breve{g}+e)}\Big)^2}{\mathrm{Var}\big(y_{\mathrm{pred}}\big)} = \frac{\mathrm{Cov}\big(y_{\mathrm{pred}},\breve{g}\big)^2}{\mathrm{Var}\big(y_{\mathrm{pred}}\big)\left[\mathrm{Var}(\breve{g})+\mathrm{Var}(e)\right]} \\
&= \left(\frac{\mathrm{Cov}\big(y_{\mathrm{pred}},\breve{g}\big)^2}{\mathrm{Var}\big(y_{\mathrm{pred}}\big)\ \mathrm{Var}(\breve{g})}\right)\left(\frac{\mathrm{Var}(\breve{g})}{\mathrm{Var}(\breve{g})+\mathrm{Var}(e)}\right) \\
&= \left(\frac{\mathrm{Var}\big(\breve{g}_{\mathrm{pred}}\big)}{\mathrm{Var}\big(y_{\mathrm{pred}}\big)}\cdot\frac{\mathrm{Cov}\big(y_{\mathrm{pred}},\breve{g}\big)^2}{\mathrm{Var}\big(\breve{g}_{\mathrm{pred}}\big)\ \mathrm{Var}(\breve{g})}\right)\left(\frac{\breve{h}^2_{\mathrm{GWAS}}}{\breve{h}^2_{\mathrm{GWAS}}+\mathrm{Var}(e)/\,\mathrm{Var}\big(y_{\mathrm{GWAS}}\big)}\right) \\
&= \left(\breve{h}^2_{\mathrm{pred}}\,r^2_{\mathbf{x},\boldsymbol{\beta}}\right)\left(\frac{\breve{h}^2_{\mathrm{GWAS}}}{\breve{h}^2_{\mathrm{GWAS}}+M/N}\right).
\end{aligned}
$$

where M is a constant and N is the GWAS sample size underlying the PGI weights; the derivation relies on the approximation $\mathrm{Cov}\,(y,e)=0$ discussed in main text Section IV.A. The second-to-last equality follows from the definition of the optimal predictor, which implies $\mathrm{Cov}\,(y_{pred},\breve{g})=\mathrm{Cov}\,(\breve{g}_{pred}+\varepsilon_{pred},\breve{g})=\mathrm{Cov}\,(\breve{g}_{pred},\breve{g})$, where ε_{pred} is the error in predicting y_{pred} in the prediction population using $\breve{g}_{pred}$. The last equality follows because $\mathrm{Var}\,(e)$ converges to zero with the GWAS sample size at rate $1/N$.

In what follows, we relax the assumption that the GWAS and prediction samples have a common LD matrix. Wang et al. (2020) and Ding et al. (2023) also relaxed **this assumption** but did so in a random-effects framework. Hence, their derivations are valid given their parametric assumptions on the joint distribution of effect sizes across the two samples. Like us, Wientjes et al. (2015, 2016) relaxed **the assumption** without making parametric assumptions but do not formally define and interpret all the parameters.[26]

We begin by establishing some notation. First, let

$$y_{pred}=\tilde{\mathbf{x}}_{pred}\breve{\boldsymbol{\beta}}_{pred}+\tilde{\varepsilon}_{pred}.$$

Here, y_{pred} is the phenotype in the prediction population, $\tilde{\mathbf{x}}_{pred}$ is the vector of observed SNP genotypes, $\breve{\boldsymbol{\beta}}_{pred}$ is the vector of optimal predictor weights in the prediction population, and $\tilde{\varepsilon}_{pred}$ is a residual that is uncorrelated with the

[26]For example, Wientjes et al. (2015; 2016) introduce a term that they call the "genetic correlation between populations," but that object is never clearly defined, and does not correspond to any of objects in our framework.

genotypes. Next, let

$$\hat{g} = \frac{\tilde{\mathbf{x}}_{\text{pred}}\,\hat{\boldsymbol{\beta}}^{\text{GWAS}}}{\text{sd}\big(\tilde{\mathbf{x}}_{\text{pred}}\,\hat{\boldsymbol{\beta}}^{\text{GWAS}}\big)} = \frac{\tilde{\mathbf{x}}_{\text{pred}}\,\big(\hat{\boldsymbol{\beta}}^{\text{GWAS}} + \mathbf{u}_{\text{GWAS}}\big)}{\text{sd}\big(\tilde{\mathbf{x}}_{\text{pred}}\,\hat{\boldsymbol{\beta}}^{\text{GWAS}}\big)}.$$

denote a PGI constructed in the prediction population using estimates of PGI weights from the GWAS population, $\hat{\boldsymbol{\beta}}^{\text{GWAS}}$, and let $\mathbf{u}_{GWAS}$ denote the estimation error from such a projection in a finite sample. Finally, let $\boldsymbol{\Sigma}_{pred} \equiv \text{Var}\left(\tilde{\mathbf{x}}_{pred}\right)$ and $\boldsymbol{\Sigma}_{GWAS} \equiv \text{Var}\left(\tilde{\mathbf{x}}_{GWAS}\right)$ denote the LD matrices in the prediction and GWAS populations, respectively. Using this notation, the R^2 from a regression of the phenotype on the PGI in the prediction sample is:

$$\begin{aligned}
R^2 &= \frac{\text{Cov}\big(y_{\text{pred}}, \hat{g}\big)^2}{\text{Var}\big(y_{\text{pred}}\big)\ \text{Var}\big(\hat{g}\big)} \\
&= \frac{\text{Cov}\Big(\tilde{\mathbf{x}}_{\text{pred}}\breve{\boldsymbol{\beta}}_{\text{pred}} + \tilde{\varepsilon}_{\text{pred}},\ \frac{\tilde{\mathbf{x}}_{\text{pred}}\left(\hat{\boldsymbol{\beta}}^{\text{GWAS}}+\mathbf{u}_{\text{GWAS}}\right)}{\text{sd}\left(\tilde{\mathbf{x}}_{\text{pred}}\hat{\boldsymbol{\beta}}^{\text{GWAS}}\right)}\Big)^2}{\text{Var}\big(y_{\text{pred}}\big)\ \text{Var}\Big(\frac{\tilde{\mathbf{x}}_{\text{pred}}\left(\hat{\boldsymbol{\beta}}^{\text{GWAS}}+\mathbf{u}_{\text{GWAS}}\right)}{\text{sd}\left(\tilde{\mathbf{x}}_{\text{pred}}\hat{\boldsymbol{\beta}}^{\text{GWAS}}\right)}\Big)} \\
&= \frac{\text{Cov}\big(\tilde{\mathbf{x}}_{\text{pred}}\breve{\boldsymbol{\beta}}_{\text{pred}},\ \tilde{\mathbf{x}}_{\text{pred}}\hat{\boldsymbol{\beta}}^{\text{GWAS}}\big)^2}{\text{Var}\big(y_{\text{pred}}\big)\ \text{Var}\big(\tilde{\mathbf{x}}_{\text{pred}}\hat{\boldsymbol{\beta}}^{\text{GWAS}}\big)} \\
&= \underbrace{\frac{\text{Var}\big(\tilde{\mathbf{x}}_{\text{pred}}\breve{\boldsymbol{\beta}}_{\text{pred}}\big)}{\text{Var}\big(y_{\text{pred}}\big)}}_{=\,\breve{h}^2_{\text{pred}}}\ \underbrace{\frac{\text{Cov}\big(\tilde{\mathbf{x}}_{\text{pred}}\breve{\boldsymbol{\beta}}_{\text{pred}},\ \tilde{\mathbf{x}}_{\text{pred}}\hat{\boldsymbol{\beta}}^{\text{GWAS}}\big)^2}{\text{Var}\big(\tilde{\mathbf{x}}_{\text{pred}}\breve{\boldsymbol{\beta}}_{\text{pred}}\big)\ \text{Var}\big(\tilde{\mathbf{x}}_{\text{pred}}\hat{\boldsymbol{\beta}}^{\text{GWAS}}\big)}}_{=\,r_g^2}\cdot\frac{\text{Var}\big(\tilde{\mathbf{x}}_{\text{pred}}\hat{\boldsymbol{\beta}}^{\text{GWAS}}\big)}{\text{Var}\big(\tilde{\mathbf{x}}_{\text{pred}}\hat{\boldsymbol{\beta}}^{\text{GWAS}}\big)}.
\end{aligned}$$

where $\breve{h}^2_{pred}$ is the optimal predictive power in the prediction sample. Hence:

$$R^2 = \breve{h}^2_{\text{pred}}\, r_g^2\, \frac{\text{Var}\big(\tilde{\mathbf{x}}_{\text{pred}}\hat{\boldsymbol{\beta}}^{\text{GWAS}}\big)}{\text{Var}\big(\tilde{\mathbf{x}}_{\text{pred}}\hat{\boldsymbol{\beta}}^{\text{GWAS}}\big)}. \tag{13}$$

For some intuition on how to interpret the parameter r_g^2, consider first the special case when $\boldsymbol{\Sigma}_{pred} = \boldsymbol{\Sigma}_{GWAS}$, $\breve{\boldsymbol{\beta}}_{GWAS} = \tilde{\boldsymbol{\beta}}^{\text{GWAS}}$, and $\breve{\boldsymbol{\beta}}_{pred} = \tilde{\boldsymbol{\beta}}_{pred}$. Then r_g^2 is the squared correlation between two additive SNP factors, one based on the GWAS weights ($\tilde{\boldsymbol{\beta}}_{GWAS}$) and one based on the prediction sample weights ($\tilde{\boldsymbol{\beta}}_{pred}$), so r_g is an instance of the genetic correlation parameter $r_{\boldsymbol{x\beta}}$. In the more general case when $\boldsymbol{\Sigma}_{pred} \neq \boldsymbol{\Sigma}_{GWAS}$, $\breve{\beta}_{GWAS} \neq \tilde{\beta}_{GWAS}$, and $\breve{\beta}_{pred} \neq \tilde{\beta}_{pred}$, r_g^2 is the correlation between the optimal predictor in the prediction population and a PGI in the prediction population that uses the GWAS-sample optimal predictor weights. Observing that $\text{Var}\big(\tilde{\mathbf{x}}_{\text{pred}}\hat{\boldsymbol{\beta}}^{\text{GWAS}}\big) = \text{Var}\big(\tilde{\mathbf{x}}_{\text{pred}}\hat{\boldsymbol{\beta}}^{\text{GWAS}}\big) + \text{Var}\big(\tilde{\mathbf{x}}_{\text{pred}}\mathbf{u}_{\mathbf{GWAS}}\big)$, Equation (13) can be rewritten as:

$$(14) \qquad R^2 = \breve{h}^2_{\text{pred}}\, r_g^2 \left(\frac{\text{Var}\big(\tilde{\mathbf{x}}_{\text{pred}}\hat{\boldsymbol{\beta}}^{\text{GWAS}}\big)\,/\,\text{Var}\big(y_{\text{GWAS}}\big)}{\text{Var}\big(\tilde{\mathbf{x}}_{\text{pred}}\hat{\boldsymbol{\beta}}^{\text{GWAS}}\big)\,/\,\text{Var}\big(y_{\text{GWAS}}\big) + \text{Var}\big(\tilde{\mathbf{x}}_{\text{pred}}\mathbf{u}_{\text{GWAS}}\big)\,/\,\text{Var}\big(y_{\text{GWAS}}\big)} \right)$$

Next, consider the term common to the numerator and denominator.

$$\begin{aligned}
\frac{\text{Var}\big(\tilde{\mathbf{x}}_{\text{pred}}\breve{\boldsymbol{\beta}}_{\text{GWAS}}\big)}{\text{Var}\big(y_{\text{GWAS}}\big)} &= \frac{\breve{\boldsymbol{\beta}}^{\top}_{\text{GWAS}}\boldsymbol{\Sigma}_{\text{pred}}\breve{\boldsymbol{\beta}}_{\text{GWAS}}}{\text{Var}\big(y_{\text{GWAS}}\big)} \\
&= \frac{\breve{\boldsymbol{\beta}}^{\top}_{\text{GWAS}}\big(\boldsymbol{\Sigma}_{\text{pred}} - \boldsymbol{\Sigma}_{\text{GWAS}} + \boldsymbol{\Sigma}_{\text{GWAS}}\big)\breve{\boldsymbol{\beta}}_{\text{GWAS}}}{\text{Var}\big(y_{\text{GWAS}}\big)} \\
&= \breve{h}^2_{\text{GWAS}} + \frac{\breve{\boldsymbol{\beta}}^{\top}_{\text{GWAS}}\big(\boldsymbol{\Sigma}_{\text{pred}} - \boldsymbol{\Sigma}_{\text{GWAS}}\big)\breve{\boldsymbol{\beta}}_{r\text{GWAS}}}{\text{Var}\big(y_{\text{GWAS}}\big)}.
\end{aligned}$$

With $\boldsymbol{\Delta}_{\boldsymbol{\Sigma}} \equiv \boldsymbol{\Sigma}_{pred} - \boldsymbol{\Sigma}_{GWAS}$, we obtain:

$$(15) \qquad \frac{\text{Var}\big(\tilde{\mathbf{x}}_{\text{pred}}\breve{\boldsymbol{\beta}}_{\text{GWAS}}\big)}{\text{Var}\big(y_{\text{GWAS}}\big)} = \breve{h}^2_{\text{GWAS}} + \frac{\breve{\boldsymbol{\beta}}^{\top}_{\text{GWAS}}\big(\boldsymbol{\Sigma}_{\text{pred}} - \boldsymbol{\Sigma}_{\text{GWAS}}\big)\breve{\boldsymbol{\beta}}_{\text{GWAS}}}{\text{Var}\big(y_{\text{GWAS}}\big)}.$$

By the properties of least-squares projection we also have:

$$\text{Var}(\mathbf{u}_{\text{GWAS}}) \;\approx\; \frac{\text{Var}(y_{\text{GWAS}})}{N}\,\boldsymbol{\Sigma}^{-1}_{\text{GWAS}}.$$

This approximation requires that the GWAS association of each SNP be small such that $\text{Var}\left(y_{GWAS}\right)$ is approximately equal to the variance of the residual of each univariate GWAS regression and that the sample size is large enough that the GWAS estimates have converged to their asymptotic distribution. Therefore, we anticipate that this approximation will be extremely good for virtually all PGIs for complex phenotypes constructed from a large-sample GWAS. Furthermore, $\tilde{\mathbf{x}}_{pred}$ and $\mathbf{u}_{\text{GWAS}}$ are mean-zero and independent. The second term in the denominator can therefore be expressed as follows:

$$\begin{aligned}
(16) \qquad \frac{\text{Var}(\tilde{\mathbf{x}}_{\text{pred}}\mathbf{u}_{\text{GWAS}})}{\text{Var}(y_{\text{GWAS}})} &= \frac{\sum[\text{Var}(\tilde{\mathbf{x}}_{\text{pred}}) \circ \text{Var}(\mathbf{u}_{\text{GWAS}})]}{\text{Var}(y_{\text{GWAS}})} \\
&\approx \frac{\sum\left[\text{Var}(\tilde{\mathbf{x}}_{\text{pred}}) \circ \frac{\text{Var}(y_{\text{GWAS}})}{N}\,\boldsymbol{\Sigma}^{-1}_{\text{GWAS}}\right]}{\text{Var}(y_{\text{GWAS}})} \\
&= \frac{1}{N}\sum\left(\boldsymbol{\Sigma}_{\text{pred}} \circ \boldsymbol{\Sigma}^{-1}_{\text{GWAS}}\right).
\end{aligned}$$

where $\circ$ denotes the element-wise multiplication operator and $\text{sum}(\cdot)$ denotes the grand sum (i.e., the sum over all the elements of the matrix). Substituting Equations (15) and (16) into (14) and rearranging yields the generalized formula: (17)

$$R^2 = \breve{h}^2_{\text{pred}}\, r_g^2 \left(\frac{\breve{h}^2_{\text{GWAS}} + \dfrac{(\breve{\boldsymbol{\beta}}^{\text{GWAS}})' \, \boldsymbol{\Delta}_{\boldsymbol{\Sigma}} \, \breve{\boldsymbol{\beta}}^{\text{GWAS}}}{\text{Var}(y_{\text{GWAS}})}}{\breve{h}^2_{\text{GWAS}} + \dfrac{(\breve{\boldsymbol{\beta}}^{\text{GWAS}})' \, \boldsymbol{\Delta}_{\boldsymbol{\Sigma}} \, \breve{\boldsymbol{\beta}}^{\text{GWAS}}}{\text{Var}(y_{\text{GWAS}})} + \dfrac{1}{N} \sum\left(\boldsymbol{\Sigma}_{\text{pred}} \circ \boldsymbol{\Sigma}^{-1}_{\text{GWAS}}\right)} \right).$$

A Remarks on Generalized Formula

For some insight into the properties of the generalized formula, it is instructive to consider the special case where the LD matrices in the populations are both diagonal. Then Equation (17) can be expressed as:

$$\frac{\text{Var}(\tilde{\mathbf{x}}_{\text{pred}}\, \mathbf{u}_{\text{GWAS}})}{\text{Var}(y_{\text{GWAS}})} \;\approx\; \frac{1}{N} \sum_{j=1}^{M} \frac{\sigma^2_{\text{pred},j}}{\sigma^2_{\text{GWAS},j}}\,.$$

In what follows, we will treat $\sigma^2_{pred,j}$ and $\sigma^2_{GWAS,j}$ as identically distributed random variables[27] and examine two benchmark cases: one in which $\sigma^2_{pred,j}$=$\sigma^2_{GWAS,j}$ and another in which they are independent. The first would arise if the GWAS and prediction populations are the same. The second is an extreme case that may arise if the two populations had been separated for an arbitrarily long time and there are no **natural selection** forces that cause allele frequencies to be similar. Under first scenario, we obtain

$$\frac{1}{N} \sum_{j=1}^{M} \frac{\sigma^2_{pred,j}}{\sigma^2_{GWAS,j}} = \frac{1}{N} \sum_{j=1}^{M} \frac{\sigma^2_{pred,j}}{\sigma^2_{pred,j}} = \frac{1}{N} \sum_{j=1}^{M} 1 = \frac{M}{N},$$

consistent with the analytical results reported in Daetwyler, Villanueva and Woolliams (2008) and de Vlaming et al. (2017). To see this, note that if

[27] We believe this assumption is reasonable for PGIs constructed using the Bayesian methods we focus on in this paper which use all measured SNPs, as long as the main driver of allele frequency differences between the prediction and GWAS populations is genetic drift. However, this assumption is likely to be violated for PGIs that are constructed using a "pruning-and-thresholding" approach, in which only a set of approximately uncorrelated SNPs that meet some statistical-significance threshold in the GWAS are included in the PGI. Under genetic drift, even though SNP effect sizes are equal across the prediction and GWAS populations, SNP allele frequencies will randomly differ, and hence $\sigma^2_{GWAS,j}$ and $\sigma^2_{pred,j}$ will randomly differ. Because inclusion in the PGI is conditioned on statistical significance, SNPs with a high $\sigma^2_{GWAS,j}$ are more likely to be included in the PGI since those SNPs will have a smaller standard error. By regression to the mean, $\sigma^2_{GWAS,j} \geq \sigma^2_{pred,j}$ for these SNPs on average, so $\sigma^2_{GWAS,j}$ and $\sigma^2_{pred,j}$ would not be identically distributed.

$\sigma^2_{pred,j} = \sigma^2_{GWAS,j}$, then $\Delta_{\mathbf{\Sigma}}$ is a null matrix and $r_g = r_{\boldsymbol{x\beta}}$. Therefore,

$$R^2 = \breve{h}^2_{pred} r^2_{\boldsymbol{x\beta}} \left(\frac{\breve{h}^2_{GWAS}}{\breve{h}^2_{GWAS} + \frac{M}{N}} \right),$$

which is exactly Equation (7) **in the main text**. Under the second scenario, the expected value of the $\text{Var}\left(\tilde{\mathbf{x}}_{pred}\mathbf{u}_{GWAS}\right)/\text{Var}\left(y_{GWAS}\right)$ term is:

$$\begin{aligned}
\mathbb{E}\left(\frac{1}{N} \sum_{j=1}^{M} \frac{\sigma^2_{pred,j}}{\sigma^2_{GWAS,j}} \right) &= \frac{1}{N} \sum_{j=1}^{M} \mathbb{E}\left(\frac{\sigma^2_{pred,j}}{\sigma^2_{GWAS,j}} \right) \\
&= \frac{1}{N} \sum_{j=1}^{M} \mathbb{E}\left(\sigma^2_{pred,j}\right) \mathbb{E}\left(\frac{1}{\sigma^2_{GWAS,j}} \right) \\
&\geq \frac{1}{N} \sum_{j=1}^{M} \frac{\mathbb{E}\left(\sigma^2_{pred,j}\right)}{\mathbb{E}\left(\sigma^2_{GWAS,j}\right)} \\
&= \frac{1}{N} \sum_{j=1}^{M} 1 \\
&= \frac{M}{N},
\end{aligned}$$

where the inequality follows from Jensen's inequality, as $f(x) = 1/x$ is convex on $(0, \infty)$. The result implies that when the GWAS and prediction population differ in LD structure, prediction accuracy falls due to the increase in $\text{Var}\left(\tilde{\mathbf{x}}_{pred}\mathbf{u}_{GWAS}\right)/\text{Var}\left(y_{GWAS}\right)$.

III Causal Interpretation of PGI

Here, we derive the coefficients from a regression of some phenotype on the child's and parental PGIs, and we show that the coefficient on the child's PGI is a weighted sum of causal effects of the child genotypes. Similar derivations can be found in Veller, Przeworski and Coop (2024) and Veller and Coop (2024). The primary difference in the derivation below is that we directly derive the coefficient on the child's PGI in a trio design, whereas previous work directly derived the coefficient in a sibling-difference design. While the coefficient on the child's PGI is the same under both designs **under the assumption of** no sibling genetic effects (as shown by Veller, Przeworski and Coop (2024)), the trio-based derivation below requires us to directly model assortative mating but allows us to additionally derive an expression for the coefficient associated with the parental

PGI.[28]

Let $\mathbf{x}$ denote a vector of genotypes for some person, $\mathbf{x}_m$ denote the vector of genotypes for the person's mother, and $\mathbf{x}_f$ denote the vector of genotypes of the person's father. We use $\mathbf{x}_p$ to denote the sum of parental genotypes

$$\mathbf{x}_p = \mathbf{x}_m + \mathbf{x}_f.$$

We define $\mathbf{x}_p$ as the sum rather than the mean in this case because it simplifies later expressions in this derivation.

To begin, we evaluate the variance-covariance (VCV) matrices for the genotype vectors. No steady-state or equilibrium restriction is imposed on the system; in particular, the VCV matrices need not be identical across generations. We split each of the genotype vectors (with elements in $\{0, 1, 2\}$) into the sum of two vectors, each with elements in $\{0, 1\}$, corresponding to the alleles inherited from each parent. We use $\mathbf{x}^{(m)}$, $\mathbf{x}_m^{(m)}$, and $\mathbf{x}_f^{(m)}$ to denote the maternally inherited genotypes for the individual, their mother, and their father, respectively (and define the paternally inherited genotypes, $\mathbf{x}^{(f)}$, $\mathbf{x}_m^{(f)}$, and $\mathbf{x}_f^{(f)}$, analogously). Then:

$$\mathbf{x} = \mathbf{x}^{(m)} + \mathbf{x}^{(f)}, \quad \mathbf{x}_m = \mathbf{x}_m^{(m)} + \mathbf{x}_m^{(f)}, \quad \mathbf{x}_f = \mathbf{x}_f^{(m)} + \mathbf{x}_f^{(f)}.$$

We denote the VCV matrix of the maternally or paternally inherited alleles by:

$$\mathbf{\Sigma} \equiv \mathrm{Var}\left(\mathbf{x}_m^{(m)}\right) = \mathrm{Var}\left(\mathbf{x}_m^{(f)}\right) = \mathrm{Var}\left(\mathbf{x}_f^{(m)}\right) = \mathrm{Var}\left(\mathbf{x}_f^{(f)}\right).$$

We similarly denote the covariance between parental genotypes by:

$$\mathbf{A} \equiv \mathrm{Cov}\left(\mathbf{x}_m^{(m)}, \mathbf{x}_f^{(m)}\right) = \mathrm{Cov}\left(\mathbf{x}_m^{(m)}, \mathbf{x}_f^{(f)}\right) = \mathrm{Cov}\left(\mathbf{x}_m^{(f)}, \mathbf{x}_f^{(m)}\right) = \mathrm{Cov}\left(\mathbf{x}_m^{(f)}, \mathbf{x}_f^{(f)}\right)$$

Finally, assortative mating in the grandparental or earlier generations may have induced correlations between the maternally and paternally genotypes inherited genotypes of each individual. We denote this covariance:

$$\mathbf{B} \equiv \mathrm{Cov}\left(\mathbf{x}_m^{(m)}, \mathbf{x}_m^{(f)}\right) = \mathrm{Cov}\left(\mathbf{x}_f^{(m)}, \mathbf{x}_f^{(f)}\right)$$

[28] Veller and Coop (2024) derive the expression for the parental coefficients for a single-SNP regression. Many of the challenges to interpreting such coefficients are qualitatively similar to those we describe below in the PGI context.

Using this notation, we can express the VCV of each parent's genotype vectors.

$$\begin{aligned}\mathrm{Var}\left(\mathbf{x}_m\right) &= \mathrm{Var}\left(\mathbf{x}_m^{(m)}\right) + \mathrm{Var}\left(\mathbf{x}_m^{(f)}\right) + 2\,\mathrm{Cov}\left(\mathbf{x}_m^{(m)}, \mathbf{x}_m^{(f)}\right) \\ &= 2\boldsymbol{\Sigma} + 2\mathbf{B} \\ &= 2\left(\boldsymbol{\Sigma} + \mathbf{B}\right).\end{aligned}$$

Similarly, $\mathrm{Var}\left(\mathbf{x}_f\right) = 2\left(\boldsymbol{\Sigma} + \mathbf{B}\right)$. Next, we calculate

$$\begin{aligned}\mathrm{Cov}\left(\mathbf{x}_m, \mathbf{x}_f\right) &= \mathrm{Cov}\left(\mathbf{x}_m^{(m)}, \mathbf{x}_f^{(m)}\right) + \mathrm{Cov}\left(\mathbf{x}_m^{(m)}, \mathbf{x}_f^{(f)}\right) \\ &\quad + \mathrm{Cov}\left(\mathbf{x}_m^{(f)}, \mathbf{x}_f^{(m)}\right) + \mathrm{Cov}\left(\mathbf{x}_m^{(f)}, \mathbf{x}_f^{(f)}\right) \\ &= 4\mathbf{A}.\end{aligned}$$

Using these results,

$$\begin{aligned}\mathrm{Var}\left(\mathbf{x}_p\right) &= \mathrm{Var}\left(\mathbf{x}_m\right) + \mathrm{Var}\left(\mathbf{x}_f\right) + 2\,\mathrm{Cov}\left(\mathbf{x}_m, \mathbf{x}_f\right) \\ &= 2\left(\boldsymbol{\Sigma} + \mathbf{B}\right) + 2\left(\boldsymbol{\Sigma} + \mathbf{B}\right) + 8\mathbf{A} \\ &= 4\left[\left(\boldsymbol{\Sigma} + \mathbf{B}\right) + 2\mathbf{A}\right].\end{aligned}$$

Next, we calculate the VCV matrix for the child's genotypes. To do this, we first consider the VCV matrix for **the** maternally or paternally inherited alleles separately. Considering a pair of alleles inherited from a particular parent, if they both had been inherited from that parent's mother or both from that parent's father, those genotypes would have a covariance given by some element of the $\boldsymbol{\Sigma}$ matrix. If they had been inherited from different parents, those genotypes would have a covariance given by some element of the $\mathbf{B}$ matrix. We let $\mathbf{P}$ denote a matrix, each of whose entries is the probability that a pair of genotypes from $\mathbf{x}^{(m)}$ are drawn from the same grandparent; this same matrix also encodes the probability that a pair of genotypes from $\mathbf{x}^{(f)}$ are drawn from the same grandparent. By the laws of Mendelian inheritance, $\mathbf{P}$ will have values of one along the diagonal, will have values of $1/2$ for any pair of genotypes corresponding to different chromosomes, and will have values between $1/2$ and one for genotypes on the same chromosome. Since the means of $\mathbf{x}^{(m)}$ and $\mathbf{x}^{(f)}$ are the same no matter which grandparent they are inherited from,

$$\mathrm{Var}\left(\mathbf{x}^{(m)}\right) = \mathrm{Var}\left(\mathbf{x}^{(f)}\right) = \mathbf{P} \circ \boldsymbol{\Sigma} + (\mathbf{1} - \mathbf{P}) \circ \mathbf{B},$$

where $\circ$ is element-wise multiplication and $\mathbf{1}$ is a matrix all of whose entries are

one. Therefore,

$$\begin{aligned}\mathrm{Var}\left(\mathbf{x}\right) &= \mathrm{Var}\left(\mathbf{x}^{(m)}\right) + \mathrm{Var}\left(\mathbf{x}^{(f)}\right) + 2\,\mathrm{Cov}\left(\mathbf{x}^{(m)}, \mathbf{x}^{(f)}\right) \\ &= 2\left[\mathbf{P} \circ \boldsymbol{\Sigma} + (\mathbf{1} - \mathbf{P}) \circ \mathbf{B} + \mathbf{A}\right].\end{aligned}$$

Finally, we calculate the covariance between the child's and parental genotype vectors:

$$\begin{aligned}\mathrm{Cov}\left(\mathbf{x}, \mathbf{x}_p\right) &= \mathrm{Cov}\left(\mathbf{x}^{(m)} + \mathbf{x}^{(f)}, \mathbf{x}_m + \mathbf{x}_f\right) \\ &= \mathrm{Cov}\left(\mathbf{x}^{(m)}, \mathbf{x}_m\right) + \mathrm{Cov}\left(\mathbf{x}^{(f)}, \mathbf{x}_f\right) \\ &+ \mathrm{Cov}\left(\mathbf{x}^{(m)}, \mathbf{x}_f\right) + \mathrm{Cov}\left(\mathbf{x}^{(f)}, \mathbf{x}_m\right) \\ &= 2\left[(\boldsymbol{\Sigma} + \mathbf{B}) + 2\mathbf{A}\right].\end{aligned}$$

Projecting y onto $\mathbf{x}$ and $\mathbf{x}_p$, we obtain the following regression equation:

$$y = \mathbf{x}\boldsymbol{\beta} + \mathbf{x}_p\mathbf{b}_p + \xi,$$

where ξ is the residual. Because of the random assignment of genotypes conditional on parental genotypes, the entries of $\boldsymbol{\beta}$ are (local average) causal effects of each genotype on y (see Appendix I). The vector $\mathbf{b}_p$ must then pick up, in addition to parental genetic effects, any gene-environment correlations (including population stratification).

Suppose we construct a PGI with weight vector $\mathbf{w}$. The PGI of the individual is $\mathbf{xw}$, and the parental PGI is $\mathbf{x}_p\mathbf{w}$. We will show that when we regress y on $\mathbf{xw}$ and $\mathbf{x}_p\mathbf{w}$, the coefficient associated with $\mathbf{xw}$ will only be a weighted sum of the elements of the causal effect vector $\boldsymbol{\beta}$ and not a function of $\mathbf{b}_p$.

Let $\alpha = [\alpha_g; \alpha_p]$ denote the population coefficients from regressing y onto $\mathbf{xw}$ and

$\mathbf{x}_p\mathbf{w}$. We calculate:

$$
\begin{aligned}
\alpha &= \begin{bmatrix} \operatorname{Var}(\mathbf{xw}) & \operatorname{Cov}(\mathbf{xw}, \mathbf{x}_p\mathbf{w}) \\ & \operatorname{Var}(\mathbf{x}_p\mathbf{w}) \end{bmatrix}^{-1} \begin{bmatrix} \operatorname{Cov}(\mathbf{xw}, y) \\ \operatorname{Cov}(\mathbf{x}_p\mathbf{w}, y) \end{bmatrix} \\
&= \begin{bmatrix} 2\mathbf{w}'\left[\mathbf{P}\circ\boldsymbol{\Sigma} + (\mathbf{1}-\mathbf{P})\circ\mathbf{B} + \mathbf{A}\right]\mathbf{w} & 2\mathbf{w}'\left[(\boldsymbol{\Sigma}+\mathbf{B}) + 2\mathbf{A}\right]\mathbf{w} \\ & 4\mathbf{w}'\left[(\boldsymbol{\Sigma}+\mathbf{B}) + 2\mathbf{A}\right]\mathbf{w} \end{bmatrix}^{-1} \\
&\times \begin{bmatrix} \operatorname{Cov}(\mathbf{xw}, \mathbf{x}\boldsymbol{\beta} + \mathbf{x}_p\mathbf{b}_p + e) \\ \operatorname{Cov}(\mathbf{x}_p\mathbf{w}, \mathbf{x}\boldsymbol{\beta} + \mathbf{x}_p\mathbf{b}_p + e) \end{bmatrix} \\
&= \begin{bmatrix} 2\mathbf{w}'\left[\mathbf{P}\circ\boldsymbol{\Sigma} + (\mathbf{1}-\mathbf{P})\circ\mathbf{B} + \mathbf{A}\right]\mathbf{w} & 2\mathbf{w}'\left[(\boldsymbol{\Sigma}+\mathbf{B}) + 2\mathbf{A}\right]\mathbf{w} \\ & 4\mathbf{w}'\left[(\boldsymbol{\Sigma}+\mathbf{B}) + 2\mathbf{A}\right]\mathbf{w} \end{bmatrix}^{-1} \\
&\times \begin{bmatrix} 2\mathbf{w}'\left[\mathbf{P}\circ\boldsymbol{\Sigma} + (\mathbf{1}-\mathbf{P})\circ\mathbf{B} + \mathbf{A}\right]\boldsymbol{\beta} + 2\mathbf{w}'\left(\boldsymbol{\Sigma}+\mathbf{B}+2\mathbf{A}\right)\mathbf{b}_p \\ 2\mathbf{w}'\left[\boldsymbol{\Sigma}+\mathbf{B}+2\mathbf{A}\right]\boldsymbol{\beta} + 4\mathbf{w}'\left(\boldsymbol{\Sigma}+\mathbf{B}+2\mathbf{A}\right)\mathbf{b}_p \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{w}'\left[\mathbf{P}\circ\boldsymbol{\Sigma} + (\mathbf{1}-\mathbf{P})\circ\mathbf{B} + \mathbf{A}\right]\mathbf{w} & \mathbf{w}'\left[(\boldsymbol{\Sigma}+\mathbf{B}) + 2\mathbf{A}\right]\mathbf{w} \\ & 2\mathbf{w}'\left[(\boldsymbol{\Sigma}+\mathbf{B}) + 2\mathbf{A}\right]\mathbf{w} \end{bmatrix}^{-1} \\
&\times \begin{bmatrix} \mathbf{w}'\left[\mathbf{P}\circ\boldsymbol{\Sigma} + (\mathbf{1}-\mathbf{P})\circ\mathbf{B} + \mathbf{A}\right]\boldsymbol{\beta} + \mathbf{w}'\left(\boldsymbol{\Sigma}+\mathbf{B}+2\mathbf{A}\right)\mathbf{b}_p \\ \mathbf{w}'\left[\boldsymbol{\Sigma}+\mathbf{B}+2\mathbf{A}\right]\boldsymbol{\beta} + 2\mathbf{w}'\left(\boldsymbol{\Sigma}+\mathbf{B}+2\mathbf{A}\right)\mathbf{b}_p \end{bmatrix} \\
&= \frac{1}{D}\begin{bmatrix} 2\mathbf{w}'\left[(\boldsymbol{\Sigma}+\mathbf{B}) + 2\mathbf{A}\right]\mathbf{w} & -\mathbf{w}'\left[(\boldsymbol{\Sigma}+\mathbf{B}) + 2\mathbf{A}\right]\mathbf{w} \\ & \mathbf{w}'\left[\mathbf{P}\circ\boldsymbol{\Sigma} + (1-\mathbf{P})\circ\mathbf{B} + \mathbf{A}\right]\mathbf{w} \end{bmatrix} \\
&\times \begin{bmatrix} \mathbf{w}'\left[\mathbf{P}\circ\boldsymbol{\Sigma} + (\mathbf{1}-\mathbf{P})\circ\mathbf{B} + \mathbf{A}\right]\boldsymbol{\beta} + \mathbf{w}'\left(\boldsymbol{\Sigma}+\mathbf{B}+2\mathbf{A}\right)\mathbf{b}_p \\ \mathbf{w}'\left[\boldsymbol{\Sigma}+\mathbf{B}+2\mathbf{A}\right]\boldsymbol{\beta} + 2\mathbf{w}'\left(\boldsymbol{\Sigma}+\mathbf{B}+2\mathbf{A}\right)\mathbf{b}_p \end{bmatrix}
\end{aligned}
$$

where, after simplifying,

$$D = \mathbf{w}'\left[(2\mathbf{P}-\mathbf{1})\circ(\boldsymbol{\Sigma}-\mathbf{B})\right]\mathbf{w}\mathbf{w}'\left(\boldsymbol{\Sigma}+\mathbf{B}+2\mathbf{A}\right)\mathbf{w}$$

is the determinant of the inverted matrix in the first line of the above derivation. Thus, for the child-PGI coefficient, we obtain

$$
\begin{aligned}
\alpha_g &= \frac{1}{D}\mathbf{w}'\left(\boldsymbol{\Sigma}+\mathbf{B}+2\mathbf{A}\right)\mathbf{w}\mathbf{w}'\left[(2\mathbf{P}-\mathbf{1})\circ(\boldsymbol{\Sigma}-\mathbf{B})\right]\boldsymbol{\beta} \\
&= \frac{\mathbf{w}'\left(\boldsymbol{\Sigma}+\mathbf{B}+2\mathbf{A}\right)\mathbf{w}\mathbf{w}'\left[(2\mathbf{P}-\mathbf{1})\circ(\boldsymbol{\Sigma}-\mathbf{B})\right]\boldsymbol{\beta}}{\mathbf{w}'\left[(2\mathbf{P}-\mathbf{1})\circ(\boldsymbol{\Sigma}-\mathbf{B})\right]\mathbf{w}\mathbf{w}'\left(\boldsymbol{\Sigma}+\mathbf{B}+2\mathbf{A}\right)\mathbf{w}} \\
&= \left(\mathbf{w}'\left[(2\mathbf{P}-\mathbf{1})\circ(\boldsymbol{\Sigma}-\mathbf{B})\right]\mathbf{w}\right)^{-1}\mathbf{w}'\left[(2\mathbf{P}-\mathbf{1})\circ(\boldsymbol{\Sigma}-\mathbf{B})\right]\boldsymbol{\beta},
\end{aligned}
$$

which is a weighted sum of the true causal effects $\boldsymbol{\beta}$. More specifically, it is the coefficient from a generalized least squares (GLS) regression of the true effect sizes onto the PGI weights, with dispersion matrix $(2\mathbf{P}-\mathbf{1})\circ(\boldsymbol{\Sigma}-\mathbf{B})$. The observation that α_g is a weighted sum of causal effects was noted in Veller and Coop (2024), but to the best of our knowledge, the observation that the coefficient is equivalent to the GLS regression of the causal effects $\boldsymbol{\beta}$ on the PGI weights $\mathbf{w}$ has not been

made before.

For the parental PGI coefficient, we obtain

$$\begin{aligned}\alpha_p = \frac{1}{D}\Big\{&\mathbf{w}'\left[(2\mathbf{P}-\mathbf{1})\circ(\boldsymbol{\Sigma}-\mathbf{B})\right]\mathbf{w}\mathbf{w}'\left(\boldsymbol{\Sigma}+\mathbf{B}+2\mathbf{A}\right)\mathbf{b}_p \\ &+\mathbf{w}'\left[\mathbf{P}\circ\boldsymbol{\Sigma}+(\mathbf{1}-\mathbf{P})\circ\mathbf{B}+\mathbf{A}\right]\left(\mathbf{w}\boldsymbol{\beta}'-\boldsymbol{\beta}\mathbf{w}'\right)\left(\boldsymbol{\Sigma}+\mathbf{B}+2\mathbf{A}\right)\mathbf{w}\Big\} \\ &=\mathbf{w}'\left(\boldsymbol{\Sigma}+\mathbf{B}+2\mathbf{A}\right)\mathbf{w}\mathbf{w}'\left(\boldsymbol{\Sigma}+\mathbf{B}+2\mathbf{A}\right)\mathbf{b}_p \\ &+\frac{\mathbf{w}'\left[\mathbf{P}\circ\boldsymbol{\Sigma}+(\mathbf{1}-\mathbf{P})\circ\mathbf{B}+\mathbf{A}\right]\left(\mathbf{w}\boldsymbol{\beta}'-\boldsymbol{\beta}\mathbf{w}'\right)\left(\boldsymbol{\Sigma}+\mathbf{B}+2\mathbf{A}\right)\mathbf{w}}{\mathbf{w}'\left[(2\mathbf{P}-\mathbf{1})\circ(\boldsymbol{\Sigma}-\mathbf{B})\right]\mathbf{w}\mathbf{w}'\left(\boldsymbol{\Sigma}+\mathbf{B}+2\mathbf{A}\right)\mathbf{w}},\end{aligned}$$

which is a function of both $\boldsymbol{\beta}$ and $\mathbf{b}_p$.

We next consider two special cases.

First, suppose that we use the true genetic effects on y as the PGI weights (and all genetic variants with causal effects on y are included in the PGI), such that $\mathbf{w}=\boldsymbol{\beta}$. In this case,

$$\alpha_g = \left(\boldsymbol{\beta}'\left[(2\mathbf{P}-\mathbf{1})\circ(\boldsymbol{\Sigma}-\mathbf{B})\right]\boldsymbol{\beta}\right)^{-1}\boldsymbol{\beta}'\left[(2\mathbf{P}-\mathbf{1})\circ(\boldsymbol{\Sigma}-\mathbf{B})\right]\boldsymbol{\beta}=1.$$

Also $(\mathbf{w}\boldsymbol{\beta}'-\boldsymbol{\beta}\mathbf{w}')=\mathbf{0}$, so

$$\alpha_p = \left[\boldsymbol{\beta}'\left(\boldsymbol{\Sigma}+\mathbf{B}+2\mathbf{A}\right)\boldsymbol{\beta}\right]^{-1}\boldsymbol{\beta}'\left(\boldsymbol{\Sigma}+\mathbf{B}+2\mathbf{A}\right)\mathbf{b}_p,$$

which means that the coefficient on the parental PGI is simply the coefficient from a GLS regression (with dispersion matrix $\boldsymbol{\Sigma}+\mathbf{B}+2\mathbf{A}$) of the parental coefficients onto the causal genetic effects.

Second, suppose that there is no assortative mating, such that $\mathbf{A}=\mathbf{B}=\mathbf{0}$. Then:

$$\alpha_g = \left(\mathbf{w}'\left[(2\mathbf{P}-\mathbf{1})\circ\boldsymbol{\Sigma}\right]\mathbf{w}\right)^{-1}\mathbf{w}'\left[(2\mathbf{P}-\mathbf{1})\circ\boldsymbol{\Sigma}\right]\beta.$$

Recall that $P_{ij}=1/2$, implying that $2P_{ij}-1=0$ for each pair of SNPs, i and j, that are on different chromosomes. Within a chromosome, if there is random mating, then Σ_{ij} decays much more quickly than P_{ij} with distance between the SNPs. This is because the matrix $\mathbf{P}$ is approximately fixed across generations since it is related to the probability that an odd number of recombinations events will have occurred between a pair of SNPs in the genome. In contrast, the elements of $\boldsymbol{\Sigma}$ will decay by a factor of $\mathbf{P}$ in each generation of random mating. (This is because there is a P_{ij} chance that the ij correlation within a parent will be broken by recombination events in each generation.) Thus, we expect the approximation

$$(2\mathbf{P}-\mathbf{1})\circ\boldsymbol{\Sigma}\approx\boldsymbol{\Sigma}$$

to be very accurate. This gives us the expressions in the main text,

$$\alpha_g = \left(\mathbf{w}'\mathbf{\Sigma}\mathbf{w}\right)^{-1}\mathbf{w}'\mathbf{\Sigma}\boldsymbol{\beta}$$
$$\alpha_p = \left(\mathbf{w}'\mathbf{\Sigma}\mathbf{w}\right)^{-1}\mathbf{w}'\mathbf{\Sigma}\mathbf{b}_p,$$

where each coefficient has a GLS interpretation.

IV Gains In Predictive Power From PGIs as Controls

Following Rietveld et al. (2013), we calculate the gains in effective sample size that could be obtained by controlling for PGIs in a simple RCT with a treatment group and a control group. Let N_{X} denote the number of experimental participants, a proportion p of whom are assigned to the treatment. The treatment effect, τ, is estimated by running a regression:

$$y = \alpha + \sum_{j=1}^{J} \beta_j X_j + \tau I + \varepsilon,$$

where y is some outcome of interest with variance σ^2, the X_j's are the values of J baseline (non-genetic) control variables whose values were determined before the intervention, I is an indicator variable equal to 1 for subjects in the treatment group and 0 for subjects in the control group, and ε is a mean-zero error term. Due to random assignment, the treatment-effect coefficient is an unbiased estimate of the treatment effect irrespective of the X_j's included in the regression. However, the precision of the treatment-effect estimate is increasing in the joint predictive power of the X_j's. In particular, under the assumption that τ^2 is small relative σ^2 (so that the R^2 from the regression on the X_j's and the intervention I is approximately equal to the R^2 from the regression on just the X_j's), the standard error for the estimate of τ will be approximately

$$\sqrt{\frac{\sigma^2}{p(1-p)N_{\mathrm{X}}}(1-R^2_{\mathrm{X}})},$$

where R^2_{X} is the fraction of variance explained in a regression of y on the X_j's.

To quantify the value of the PGI, Rietveld et al. (2013) consider a hypothetical researcher who wishes to maximize statistical power and must choose between two alternatives:

1) Conduct the study among N_{X} participants for whom X_j's have been measured.
2) Conduct the study among $N_{\mathrm{X}\cup\mathrm{PGI}} < N_{\mathrm{X}}$ participants for whom the X_j's and a PGI have been measured, thus increasing the joint predictive power of the covariates from R^2_{X} to $R^2_{\mathrm{X}}+R^2_{\mathrm{PGI|X}}$.

The two study designs are identically powered when their expected standard errors are identical. Rietveld et al. (2013) show that which option has lower expected standard error is determined by the inequality:

$$\frac{N_{\mathrm{X}\cup\mathrm{PGI}}}{N_{\mathrm{X}}} \underset{>}{\leq} \frac{1-\left(R^2_{\mathrm{X}}+R^2_{\mathrm{PGI|X}}\right)}{1-R^2_{\mathrm{X}}}. \tag{18}$$

The left-hand side represents the proportional loss in power (in units of squared standard error) that comes from reducing the sample size from N_{X} to $N_{\mathrm{X}\cup\mathrm{PGI}}$. The right-hand side represents the proportional gain in power (in the same units) that comes from adding the PGI to the set of control variables. Adding the PGI to the set of controls generates a net gain in power if the right-hand side is larger, a net loss if it is smaller, and no change in power if the two sides are equal.

To quantify the gains in power from including a PGI with predictive power $R^2_{\mathrm{PGI|X}}$, Rietveld et al. (2013) calculate the reduction in original N_{X} that would hold power constant between the options for some given value of R^2_{X}. They then calculate gains for values of $R^2_{\mathrm{PGI|X}}$ between 2% (the amount of predictive power attainable at the time of the study) to 15%. Assuming $R^2_{\mathrm{X}} \in \{10\%, 20\%\}$, their analyses show a PGS with $R^2_{\mathrm{PGI|X}} = 15\%$ would yield benefits equivalent to reducing N_X by 17-19%.

Equation (18) shows two quantities determine the gains in effective sample size from controlling for a PGI. First, the gains are increasing in $R^2_{\mathrm{PGI|X}}$. The important nuance is that the predictive power of the PGI is not *per se* **what** matters; rather, **what matters is** its incremental predictive power over the controls that are already included. Second, the gains are increasing in R^2_X. Thus, perhaps counter-intuitively, in situations where the incremental predictive power of the PGI is small because already-included controls are highly predictive, including the PGI as a control can nonetheless add a lot in terms of effective sample size. For example, suppose the outcome is test scores and pre-treatment test scores are already included as controls with $R^2_{\mathrm{X}} = 80\%$. Then, even if $R^2_{\mathrm{PGI|X}}$ is only 2%, the benefit of including the PGI as a control is equivalent to increasing the sample size by 11%.