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 the editors, David Romer and Steven Durlauf, and four anonymous referees, as well as 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, Dhruva Jaishankar, Giorgia Mezzetti and Moeen Nehzati for research assistance. For financial support, we thank the NIA/NIH (grants R24-AG065184, R01-AG042568, R00-AG062787, R01-CA313942, R01-AG083379 and R01-AG081518) and Open Philanthropy.

1. *Introduction*

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; Björklund, Jäntti and Solon, 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, in order to obtain sufficiently large samples, early GWAS studies were conducted almost exclusively in samples of unrelated individuals, complicating causal interpretation. 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 correlational analyses, aims to shed light on 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 causal 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 mating shape genetic

[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.

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 2 begins with a rudimentary genetics primer. Section 3 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 4. Section 5 defines, interprets, and analyzes PGIs, the centerpiece of most applications in recent years. In Section 6, 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 7, 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 intrinsic to research at the intersection of genetics and social science.

[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 fourth (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 (for example, a recent textbook by Mills, Barban and Tropf, 2020).

2. *Genetics Primer*

This section provides genetics background for the rest of the paper and may be skipped and consulted as needed. To keep the exposition compact, we omit many nuances[3]; readers seeking more detail should consult textbooks in molecular genetics Strachan and Read (2018) and population genetics Gillespie (2004).

2.1. *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 position in the genome is defined by a location on a specific chromosome. 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.

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 common convention, a variant is classified as *rare* if at least 99% of individuals carry the same version

[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 2.2. We also ignore mutations that cause small differences in DNA across cells.

and as *common*) otherwise. Common variants are also called *polymorphisms*. 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), say 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 an observed 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*.

2.2. *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*.

2.3. *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. (Such very high LD is one reason why regressions on many genetic variants at once typically face a short-rank problem—a problem that partly motivates genetic study designs, as we will discuss in Section 4.)

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-increasing alleles, including those located on different chromosomes. Another, related form of assortative mating is *population structure*: individuals within a subpopulation—for example, geographic region, or with a shared ethnicity or language—tend to mate with each other (see, for example, Bergstrom, 2013). In that case, alleles that are more common within the subpopulation will become correlated, regardless of their position in the genome.

2.4. *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

[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).

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.

2.5. *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 4.1) have been from SNP arrays because SNP array genotyping has been much less expensive. Today, such genotyping costs roughly $30 per participant.

3. *Statistical Framework: Genetic Effects*

In this section, we lay out a framework for understanding and interpreting the relationship between genetic variants and complex phenotypes. Our approach builds on classic treatments in Fisher (1918) and Falconer (1960) but is expressed and interpreted using the modern, potential-outcomes conceptual apparatus introduced by Rubin (1974). Our hope is that the formal treatment helps make precise the meaning of key theoretical constructs and estimands, thereby clarifying the link between theory and empirical practice, the compromises empirical work inevitably entails, and the interpretational limits those compromises impose.

3.1. *The General Framework*

Consider a large population of individuals indexed by $i = 1, 2, ..., I$. Let y_i denote i's scalar-valued phenotype, and let

$$\mathbf{x}_i = \left(x_{i1}, x_{i2}, \ldots, x_{iJ}\right) \in \mathbb{R}^J,$$

denote i's $1 \times J$ genotype vector, where $x_{ij} \in \{0, 1, 2\}$ (before recentering) represents the number of minor alleles at variant $j = 1, 2, ..., J$. Without loss of generality, we recenter y_i and each x_{ij} so they have mean zero across individuals in the population. To distinguish them from counterfactuals, we refer to y_i and $\mathbf{x}_i$ as i's *observed* phenotype and genotype vector. We denote by $\mathbf{X} = \left(X_1, X_2, \ldots, X_J\right)$ the random vector that reflects the distribution of observed genotype vectors, with $\mathbf{x}_i$ representing the realized vector for individual i.

We adopt a potential-outcomes framework (as in Rubin, 1974) where causal effects are defined in terms of counterfactual outcomes following hypothetical interventions. Let $y_i(\mathbf{x})$ denote i's potential outcome at any genotype vector $\mathbf{x} \in \mathcal{X} = \{0, 1, 2\}^J$. We refer to $y_i(\cdot)$ as i's "potential-outcome function," and its argument $\mathbf{x}$ may be any genotype vector, possibly distinct from anyone's observed $\mathbf{x}_i$. That is, $y_i(\mathbf{x})$ is the value of i's phenotype following a hypothetical intervention that changes i's genotype (from $\mathbf{x}_i$) to $\mathbf{x}$ at conception, with everyone else's genotype vector remaining fixed at its observed value. We can then write i's observed phenotype as $y_i = y_i(\mathbf{x}_i)$.

The individual-level causal effect of changing genotype from $\mathbf{x}$ to $\mathbf{x}^\star$ is $y_i(\mathbf{x}^\star) - y_i(\mathbf{x})$. The average causal effect is the mean of these individual-level effects:

$$\mathbb{E}\left[y_i(\mathbf{x}^\star) - y_i(\mathbf{x})\right] = \frac{1}{I}\sum_{i=1}^{I}\left[y_i(\mathbf{x}^\star) - y_i(\mathbf{x})\right]. \tag{1}$$

We highlight that the definition does *not* require effects to be constant across people or environments and that *any* difference in i's phenotype that can ultimately be attributed to the hypothetical change in genotype contributes to the causal genetic effect in individual i, irrespective of the mediating pathways (we return to this latter issue in Section 3.4).

We define the *genetic factor* at genotype vector $\mathbf{x} \in \mathcal{X}$, a scalar, as the population mean of potential-outcome functions all evaluated at the same $\mathbf{x}$:

$$G(\mathbf{x}) := \mathbb{E}\left[y_i\left(\mathbf{x}\right)\right] = \frac{1}{I}\sum_{i=1}^{I} y_i(\mathbf{x}).$$

Thus, the average causal effect can be expressed as $G(\mathbf{x}^\star) - G(\mathbf{x})$. Moreover, we can express i's potential outcome at each $\mathbf{x}$ as:

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

where $\nu_i\left(\mathbf{x}\right)$ captures all factors that cause i's outcome under assignment to genotype $\mathbf{x}$ to differ from the population mean outcome under assignment to $\mathbf{x}$. By construction, $\mathbb{E}\left[\nu_i\left(\mathbf{x}\right)\right] = \frac{1}{I}\sum_{i=1}^{I}\nu_i(\mathbf{x}) = 0$ for every $\mathbf{x}$, but the framework places no constraints on the joint distribution of $\nu_i(\mathbf{X})$ and $G(\mathbf{X})$.

3.2. *The Additive Genetic Factor*

Since $G(\cdot)$ can be an arbitrary unstructured function, theoretical and empirical research typically focuses on a linear approximation called the additive genetic factor. For any $\mathbf{x} \in \mathcal{X}$, $G(\mathbf{x})$ depends on the distribution of potential outcomes $y_i(\mathbf{x})$ in the population but *not* on the distribution of observed genotype vectors $\mathbf{x}_i$. In contrast, the additive genetic factor depends on both distributions. From here on throughout Section 3, we subscript by $\mathbf{X}$ expectations and variances taken over the distribution of genotype vectors to distinguish them from expectations taken across potential-outcome functions holding fixed the genotype vector at the same $\mathbf{x}$ for everyone.

The *additive genetic factor* $g\left(\mathbf{X}\right)$ is defined by linearly projecting $G\left(\mathbf{X}\right)$ onto $\mathbf{X}$:

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

By standard arguments (Goldberger, 1991), the solution is $\boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1}\,\mathbb{E}_{\mathbf{X}}[\mathbf{X}'G(\mathbf{X})]$, where $\boldsymbol{\Sigma}$ is the symmetric $J \times J$ variance-covariance matrix,

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

known as the *linkage disequilibrium* (LD) matrix. We assume that its columns are linearly independent so that the solution to the least-squares problem is unique. 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 (by a Central Limit Theorem).

An important subtlety is that $\boldsymbol{\beta}$ and hence $g\left(\cdot\right)$, the mean-squared-error-minimizing linear approximation to $G\left(\cdot\right)$, depends on the distribution of genotype vectors $\mathbf{X}$ in the population. In general, changes to the distribution will result in changes to the coefficient vector, even if $G(\cdot)$ has not changed. Therefore, two populations with the same function $G(\cdot)$ will generally have different $\boldsymbol{\beta}$ vectors.

The deviation of $G(\cdot)$ from the best linear approximation $g(\cdot)$ is known as the *non-additive genetic factor*: $N(\mathbf{X}) = G(\mathbf{X}) - \mathbf{X}\boldsymbol{\beta}$. By construction, $N(\mathbf{X})$ is the residual from the linear projection of $G(\mathbf{X})$ onto $\mathbf{X}$. Standard arguments therefore imply that $\mathbb{E}_{\mathbf{X}}[N(\mathbf{X})] = 0$ and $\mathrm{Cov}_{\mathbf{X}}(X_j, N(\mathbf{X})) = 0$ for every $j = 1, 2, \ldots, J$ (as well as any linear combination of elements of $\mathbf{X}$). 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, if 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*, which decomposes each individual's potential-outcome function into a linear component shared across individuals and an individual-specific residual. For every $\mathbf{x} \in \mathcal{X}$,

$$y_i(\mathbf{x}) = \mathbf{x}\boldsymbol{\beta} + \epsilon_i(\mathbf{x}), \tag{2}$$

where $\epsilon_i(\mathbf{x}) = N(\mathbf{x}) + \nu_i(\mathbf{x})$. In empirical settings, the main obstacle to identification of $\boldsymbol{\beta}$ is failure of exogeneity: $\mathrm{Cov}_{\mathbf{X}}(\mathbf{X}, \epsilon_i(\mathbf{X})) \neq 0$ (see further discussion in Section 4.2). Even though $\mathrm{Cov}_{\mathbf{X}}(\mathbf{X}, N(\mathbf{X})) = 0$ holds by construction, $\mathbf{X}$ and $\nu_i(\mathbf{X})$ could be correlated. To illustrate, suppose individuals in region A have access to better schools than those in region B, leading their potential-outcome functions and hence their $v_i(\mathbf{X})$'s to be shifted upward. Then any difference in the distribution of genotype vectors between the two regions could induce a correlation between $\mathbf{X}$ and $\epsilon_i(\mathbf{X})$.

The elements of the vector $\boldsymbol{\beta}$ can be interpreted as average causal effects: $G(\mathbf{X})$ represents average causal effects across individuals; and $g(\cdot)$, being the best linear approximation to $G(\mathbf{X})$, represents the causal effects averaged over non-linearities. This interpretation remains valid if the causal effects vary across individuals, for example, due to gene-by-environment interactions. Moreover, deviations from linearity are captured by $N(\mathbf{X})$, which can be treated as error because $\mathbb{E}_{\mathbf{X}}[N(\mathbf{X})] = 0$ and $N(\mathbf{X})$ 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 estimating $\boldsymbol{\beta}$ is presumably greater when the approximation of $G(\mathbf{X})$ by $g(\mathbf{X})$ is more accurate. Theoretical work in statistical genetics finds that the approximation is often quite accurate for complex, but not monogenic, phenotypes (Hill, Goddard and Visscher, 2008). To be more precise, the difference between $G(\mathbf{X})$ and $g(\mathbf{X})$—that is, 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 relies on not-fully-compelling identification strategies but is consistent with 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 epistasis has similarly failed to detect substantial epistatic variance, though the very large number of possible interactions limits power and yields estimates with wide 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 $G(\mathbf{X})$.

3.3. *Unobserved Variants and the Additive SNP Factor*

In empirical applications, only a strict subset K of the J elements of $\mathbf{x}$ are observed, with all $K < J$ observed elements SNPs. In such situations, the *additive SNP factor,* $\tilde{g}(\mathbf{X})$, is sometimes used to proxy for the additive genetic factor $g(\mathbf{X})$. Here, we define and interpret this object.

We assume, without loss of generality, that the SNPs in the $1 \times J$ random vector $\mathbf{X}$ are ordered with the K observed SNPs first. We can then write $\mathbf{X} = (\tilde{\mathbf{X}}, \mathbf{X_u})$, where $\tilde{\mathbf{X}} = (X_1, X_2, \ldots, X_K)$ is the $1 \times K$ subvector of observed SNPs and $\mathbf{X_u} = (X_{K+1}, X_{K+2}, \ldots, X_J)$ is the $1 \times (J-K)$ subvector of unobserved variants. The realized genotype vector of observed SNPs for individual i is denoted $\tilde{\mathbf{x}}_i = (x_{i1}, x_{i2}, \ldots, x_{iK})$.

The *additive SNP factor*, $\tilde{g}(\mathbf{X})$, is defined by linearly projecting $G(\mathbf{X})$ onto $\tilde{\mathbf{X}}$:

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

In terms of definition, the only difference between the additive genetic factor $g(\mathbf{X}) = \mathbf{X}\boldsymbol{\beta}$ and the additive SNP factor $\tilde{g}(\mathbf{X}) = \tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}}$ is that the latter is defined by a linear projection onto the subvector $\tilde{\mathbf{X}}$ rather than the full vector $\mathbf{X} = (\tilde{\mathbf{X}}, \mathbf{X_u})$. As a projection of causal genetic effects, variation in the additive SNP factor across individuals, like variation in the additive genetic factor, captures variation in causal genetic effects. However, the causal interpretation of $\boldsymbol{\beta}$ does not extend to $\tilde{\boldsymbol{\beta}}$. To see why, write the additive SNP model as:

$$y_i = \tilde{\mathbf{x}}_i\tilde{\boldsymbol{\beta}} + \tilde{\epsilon}_i \tag{3}$$

(note that, unlike Equation (2), the model cannot be expressed in terms of potential outcomes, $y_i(\mathbf{x})$, since the subvector $\tilde{\mathbf{x}}_i$ does not pin down the full vector $\mathbf{x}$). Intuitively, there is omitted-variables bias: any $\tilde{\beta}_k \in \tilde{\boldsymbol{\beta}}$ will typically capture some unknown mix of the average causal effect β_k and an omitted-variable bias term that depends on the causal effects of the unobserved variants and the variance-covariance matrix of the full genotype vector, $\boldsymbol{\Sigma}$.

3.4. *Interpretational Issues: Genetic Effects*

We now use our statistical framework to try to clarify some enduring misunderstandings about genetic effects. Our treatment closely follows Jencks (1980) and Jencks and Brown (1977). We emphasize that the issues here are purely conceptual and hence distinct from practical challenges that arise in empirical settings due to measurement problems or imperfect identification strategies.

Building on the observation, highlighted in Section 3.1, that a causal genetic effect captures *all* causal pathways, Jencks (1980) proposes a taxonomy of three broad pathways through which genes may influence outcomes.[6] 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 body mass index (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" often evokes mental associations with physical mechanisms like those described in biology textbooks for monogenic disorders. Jencks (1980) terms 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 (that is, 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 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.

[6] 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.

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.

3.5. *Genetic Factors, Heritability and Genetic Correlation*

Informally, *heritability* is described as the proportion of variation in a phenotype "attributable" to genes, whereas *genetic correlation* characterizes the degree of "genetic overlap" between two phenotypes. Here, we formally define and discuss several distinct definitions of these concepts. We use these definitions to clarify their appropriate interpretation and the challenges to their credible empirical identification.

We begin with heritability. Each of the three definitions below takes the form of a ratio of the variance in one of the genetic factors to the variance in the phenotype:

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

The variances are taken across individuals: $\mathrm{Var}\left(y_i\right) = \frac{1}{I}\sum_{i=1}^{I}\left[y_i - \mathbb{E}\left(y_i\right)\right]^2$ where $\mathbb{E}\left(y_i\right) = \frac{1}{I}\sum_{i=1}^{I} y_i$ and $\mathrm{Var}\left(F(\mathbf{x}_i)\right) = \frac{1}{I}\sum_{i=1}^{I}\left[F(\mathbf{x}_i) - \mathbb{E}\big[F(\mathbf{x}_i)\big]\right]^2$ where $\mathbb{E}\big[F(\mathbf{x}_i)\big] = \frac{1}{I}\sum_{i=1}^{I} F(\mathbf{x}_i)$. Each captures some notion of how much cross-sectional variation in the phenotype can be explained by *causal effects* of genetic variation in the population. For example, consider broad heritability. Fix some arbitrary genotype vector $\mathbf{x}$ as a benchmark for measuring causal effects. Recall from Section 3.1 that the average causal effect on individual i's phenotype from having inherited observed genotype $\mathbf{x}_i$ instead of $\mathbf{x}$ is $G(\mathbf{x}_i) - G(\mathbf{x})$. The numerator for broad heritability is the amount of phenotypic variance in the population resulting only from the causal effects of each individual having inherited their observed genotype: $\mathrm{Var}\left(G(\mathbf{x}_i) - G(\mathbf{x})\right) = \mathrm{Var}\left(G(\mathbf{x}_i)\right)$, where the choice of benchmark $\mathbf{x}$ does not matter because it is fixed across individuals. Analogously, the numerators for narrow and SNP heritability are the amount of phenotypic variance in the population predicted from the best linear approximation to genetic causal effects based on the full genotype vector or the observed SNPs, respectively. Since both the numerator of denominator in every definition of heritability can differ across populations, all should be thought of as population-specific quantities, *not* structural parameters.

To further understand heritability and challenges to its empirical identification, it is useful to consider the correlation between a genetic factor $F(\mathbf{X})$ and non-genetic factors $\nu_i(\mathbf{X})$, given the distribution of genotype vectors $\mathbf{X}$ in a population. If $F(\mathbf{X})$ and $\nu_i(\mathbf{X})$ were uncorrelated, then heritability would be equal to the R^2 from a regression of y_i on $F(\mathbf{x}_i)$. Note that if $F(\mathbf{X})$ and $\nu_i(\mathbf{X})$ were negatively correlated—which could happen if, for example, policy aimed at reducing inequality had the effect of shifting upward the potential outcomes of individuals with lower genetic factors—then $\text{Var}\,(F(\mathbf{x}_i))$ could exceed $\text{Var}\,(y_i)$.[7] In such a case, heritability could exceed one. In practice, however, we generally expect the genetic and non-genetic factors that increase a phenotype to be positively correlated (in which case heritability is bounded above by one); for example, individuals with genotypes that lead to greater educational attainment are likely to be have parents with such genotypes, who create a rearing environment that also leads to greater educational attainment. In empirical research, such so-called *gene-environment correlation* is the main obstacle to identification of heritability (see Section 4.7).

Broad heritability, H^2, summarizes the extent to which genetic variation causes variation in a phenotype in a particular population. Narrow heritability, h^2, is the key quantity for quantifying the response to selection pressures, both natural and artificial (Visscher, Hill and Wray, 2008). For example, estimates of narrow heritability are used in plant and animal breeding to predict how selective breeding will change the mean of the phenotype distribution from one generation to the next. Outside these settings, narrow heritability is relevant to studying polygenic indexes (PGIs) because, as we will discuss in Section 5, a PGI is a predictor of a phenotype constructed as a weighted sum of the genotypes of observed genetic variants. To the extent that some phenotype's h^2 remains stable over time, estimates of it can provide an (estimated) upper bound on the accuracy of *future* PGIs of that phenotype within the same population that are constructed from more genetic variants than are measured by current technologies and whose weights are more accurate estimates of the genetic variants' causal effects. 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 6.3).

SNP heritability, $\tilde{h}^2$, provides an upper bound on the accuracy of *current* PGIs for the phenotype within the same population that are based on the same K observed SNPs. Estimates of this upper bound can be useful in social-science applications, as we discuss in Section 5.5. 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

[7] To illustrate, suppose the genotype vector $\mathbf{X}$ takes only two values, $\mathbf{x}_A$ and $\mathbf{x}_B$. Half the population are Type 1, with potential outcomes $y_1\,(\mathbf{x}_A) = 0$ and $y_1\,(\mathbf{x}_B) = 10$, and half are Type 2, with $y_2\,(\mathbf{x}_A) = 10$ and $y_2\,(\mathbf{x}_B) = 20$. If all Type 1 individuals have genotype $\mathbf{x}_B$ and all Type 2 individuals have genotype $\mathbf{x}_A$, then everyone's realized phenotype is 10, so $\text{Var}\,(y_i) = 0$, even though $\text{Var}\,(G(\mathbf{x}_i)) > 0$. Specifically, since $G(\mathbf{x}_A) = 5$ and $G(\mathbf{x}_B) = 15$, $\mathbb{E}\,(G(\mathbf{x}_i)) = 10$, and $\text{Var}\,(G(\mathbf{x}_i)) = 0.5 \times (5-10)^2 + 0.5 \times (15-10)^2 = 25$.

unobserved genetic variants, and quality-control filters (that is, dropping some observed SNPs) generate such a barrier to comparability, a challenge we return to in our discussion of PGIs (Section 5.5).

To understand how the three notions relate ordinally, observe that $G(\mathbf{x})$, $g(\mathbf{x})$ and $\tilde{g}(\tilde{\mathbf{x}})$ are solutions to the problem of finding the best predictor of i's potential outcome at genotype $\mathbf{x}$, $y_i(\mathbf{x})$, with successfully more restrictive constraints imposed on the functional form. For any fixed $\mathbf{x} \in \mathcal{X}$, the unconstrained solution is $G(\mathbf{x}) = \mathbb{E}[y_i(\mathbf{x})]$. The step $G(\mathbf{x}) \to g(\mathbf{x})$ comes from imposing the constraint of a linear function: $g(\mathbf{x}) = \mathbf{x}\boldsymbol{\beta}$. Finally, the step $g(\mathbf{x}) \to \tilde{g}(\mathbf{x})$ further restricts this linear approximation to observed SNPs only: $g(\mathbf{x}) = \tilde{\mathbf{x}}\tilde{\boldsymbol{\beta}}$. 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 unobserved variants has a nonzero effect on the phenotype.

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.) Both are scalars. 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 PGIs 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 are more similar. 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. We now give two examples, but we caution that these examples—like almost all existing estimates of heritability and 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. We will return to these points in Section 4.7.

As a first example, genetic correlations can provide new evidence for evaluating theories. For example, Karlsson Linnér et al. (2019) estimate coefficient-based genetic correlations between a survey-based measure of general risk tolerance and several risky behaviors across distinct domains and find that the estimates exceed the corresponding phenotypic correlations (that is, 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 (for example, 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, but relatively uncorrelated domain-specific non-genetic effects on behavior cause the risky behaviors themselves to have weak correlatives.

As a second example, the genetic correlation estimated 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 estimate 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.

3.6. *Interpretational Issues: Heritability*

We now return once again to our statistical framework to clarify stubborn misconceptions about heritability. Many of these misconceptions are directly related to misunderstandings about genetic effects discussed in Section 3.4. Our discussion here builds on insights from Jencks (1980) and Goldberger (1979).

One major misconception is that one minus heritability is the share of phenotypic variance explained by the environment. This is wrong, partly because genes may exert some of their effects through environmental variables. 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. To the extent that genetic effects operate through endogenous environments, one minus heritability *underestimates* the share of phenotypic variance explained by the environment.

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 unobserved variants (in the case of $\tilde{\epsilon}_i$). To the extent that the residual captures some genetic effects, one minus heritability *overestimates* the share of phenotypic variance explained by the environment. 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 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 (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. Moreover, the genetic factor for some phenotype picks up not only genetically influences on that phenotype but also genetic influences on other phenotypes with which it has a non-zero (factor-based) genetic correlation. 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.

3.7. *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* (for example, 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. All of the interpretational issues discussed in Section 3.4—in particular, that these genetic effects capture all causal pathways, including environmental pathways—apply also to interpersonal genetic effects. Below we discuss two types of interpersonal genetic effects, sibling and parental, to highlight some additional nuances of interpretation that can arise for 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. Current approaches to defining and studying parental genetic effects ignore these channels, and therefore existing estimates are potentially subject to selection biases. One way economists might contribute to this literature is by developing econometric approaches that separately identify these channels. 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

[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).

this definition, parental genetic effects are understood to capture the self genetic effects on the parent's phenotypes (for example, 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 under-appreciated: 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 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 6.5). 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.

4. *Estimation of Genetic Effects*

Typically the causal effects of interest are given by the vector $\boldsymbol{\beta}$, as defined in Section 3.2. 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.

Suppose first that we have a very large sample of individuals with all J genetic variants observed 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 (2). Instead, we consider the analog of Equation (2) in terms of observables:

$$y_i = \mathbf{x}_i \boldsymbol{\beta} + \epsilon_i. \tag{4}$$

We face the usual identification challenge: the ordinary least squares estimator of $\boldsymbol{\beta}$ will be biased if $\mathrm{Cov}(\mathbf{x}_i, \epsilon_i) \neq 0$. 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 in Section 4.3. In Sections 4.4 and 4.5, we will discuss how 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 unobserved genetic variants. Second, the full-rank condition fails; hence estimation of Equation (4) 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.

4.1. *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 each SNP j, the estimating equation is:

$$y = x_j \beta_j^{GWAS} + \mathbf{z}\gamma_j + \epsilon_j, \tag{5}$$

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 observed 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

[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 GWASs (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.

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 (that is, assuming the controls $\mathbf{z}$ are sufficient for credible causal inference), β_j^{GWAS} can be expressed as a function of the elements of the vector $\boldsymbol{\beta}$ from Equation (4) using the standard formula for multivariate omitted-variables bias:

$$\beta_j^{\text{GWAS}} = \sum_{k=1}^{J} \frac{r_{jk\perp\mathbf{z}}}{r_{jj\perp\mathbf{z}}} \beta_k, \tag{6}$$

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 $\boldsymbol{\beta}$, essentially by inverting Equation (6) and regularizing the β_j's (Lloyd-Jones et al., 2019).

4.2. *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 (Hamer and Sirota, 2000) involves a poorly designed case-control study of chopstick use, where "cases" are sampled from an Asian-ancestry population and controls are sampled from a different population. Such a study might falsely attribute causal genetic effects on chopsticks use to any SNP whose allele frequency varies systematically between the groups. The attribution is of course incorrect because the association between the SNP and the outcome is spurious, arising entirely because the SNP is a marker for ancestry, which in turn is correlated with the unobserved cultural factors that determine chopstick use.

A second potential source of bias arises from interpersonal genetic effects from relatives (Section 3.7). 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 feasibility, and relative credibility for causal inference.

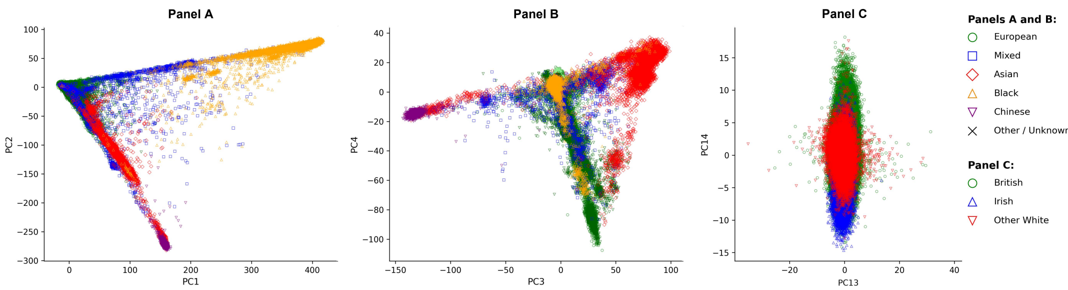
4.3. *Population-Based GWAS with Principal Components*

Most human genetic-association studies published to date have been conducted in samples of approximately unrelated individuals with a shared, continental ancestry. Recently, it has been increasingly common to label these studies *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.

Using PCs as controls has two major advantages. First, PCs can be estimated in any dataset in which a GWAS can be conducted (that is, with SNP array data). Second, controlling for PCs can mitigate confounding from population stratification because the top PCs typically capture information about geographic and ethnic ancestry (Menozzi, Piazza and Cavalli-Sforza, 1978) (and do so far better than self-reported race and ethnicity; see Rietveld et al., 2014). 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. Figure 1 provides several illustrations of how PCs can capture real genetic structure. The first two panels show that when individuals are labeled by their self-reported broad ethnic identity (Black, European, Asian, mixed, etc), several discernible clusters emerge when the top four PCs are plotted. 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. Panel C in Figure 1 illustrates that among those who identify as European, PC13 and PC14 are quite effective at discriminating between those who self-identify as British, Irish and "Other White", respectively (see also Bycroft et al., 2018).

Based on similar 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

Figure 1. Principal Components and Self-Reported Ethnic Background



Note: The principal components depicted were calculated by Bycroft et al. (2018) for all UK Biobank participants with non-missing self-reported ethnic background (N = 487,382). For the six major broad ethnic group, Panel A plots PC1 versus PC2; Panel B plots PC3 versus PC4. Each point is an individual; marker color and shape denote the individual's ethnic background (at a higher level of aggregation). Panel C is restricted to respondents who self-identify as European and plots PC13 against PC14 separately for three subcategories (British, Irish, Any Other White).

very effective way to guard against population-stratification bias, 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 (for example, 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 (for example, 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 stratification. Third, genetic PCs computed in the standard way do not capture recent population structure (that is, 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 (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.

4.4. *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:

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

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 (5) is run controlling for $x_{p,j}$,[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 A). Therefore, as Veller, Przeworski and Coop (2024) show, 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 (Imbens, 2010) emphasizing the value of a clearly defined causal estimand, recoverable under weak assumptions, whereas critics (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). However, 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

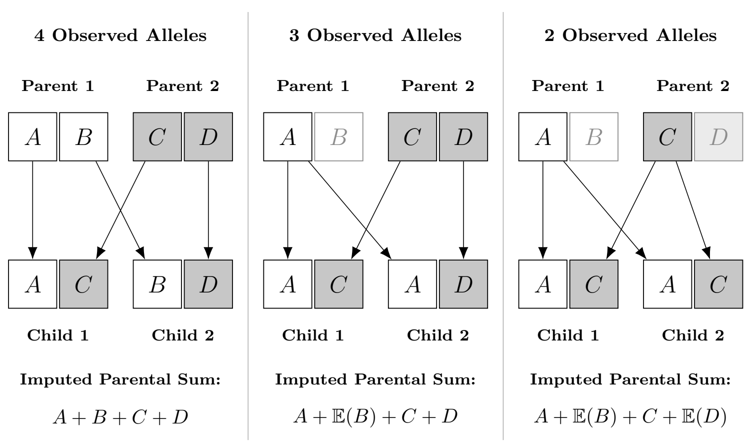
[10] A subtlety here is that $x_{r,j}$ is not independent of, but only mean independent of, $x_{p,j}$ and ϵ (the residual from Equation (4)). That is because at any genetic variant j, only heterozygous parents contribute to variation in the offspring genotype. Therefore, the variance of $x_{r,j}$ and 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 (5) 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).

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] Here, because it is also true of LATE estimates more generally, we emphasize again that the genetic effect estimate captures *all* channels through which the genotype might ultimately affect the phenotype.

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. Lightly shaded alleles (B in the middle panel and B and D in the right panel) are unobserved.

figure

The main challenge for family-based GWASs so far has been the small number of samples available for analyses that have genotyped trios (both parents and their offspring). Nonetheless, the scope for applying this identification strategy

[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 restrict the sample to have relatively homogeneous genetic ancestry (or separately analyze the sample by genetic ancestry), as mentioned in Section 4.3.

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) show 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 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).

4.5. *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 approach is equivalent to the sibling-difference estimator that regresses the sibling 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 that, despite the popularity of sibling-based approaches, are not widely appreciated (see Fletcher et al., 2024, for additional limitations). The first, described 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 (that is, the genes of one sibling do not impact the outcome of the other sibling; see Section 3.7). To see how the potential for bias arises, note that the identifying variation from the first sibling

[13] Hwang et al. (2020) propose 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.

[14] Trejo and Kanopka (2024) develop a related, alternative approach that can be applied when phenotype data is available for both siblings but genotype data is available for only one in each pair, but it requires an assumption about the magnitude of genetic assortative mating.

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 (5). 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, $x_{r,j}$. Because each sibling's genotype has additional, independent variation relative to the parental midpoint, 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.

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]

4.6. *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 (see, for example, 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 4.7 below. For both medical and social-science applications, a polygenic index is among the most important outputs of a GWAS; we defer discussion of polygenic indexes to Section 5.

[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).

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 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 are 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} are smaller in magnitude on average than the population-based GWAS estimates. Tan et al. (2024) estimate 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—have a large estimate of genetic correlation with educational attainment using population-based GWAS results, but the family-based results are 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

[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 a 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.

Section 4.3, 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 September 2025, 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 $\sim$90%, compared to their $\sim$15% share of the global population.

This "eurocentric bias" is widely considered to be a major problem (for example, 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 5.3, 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 countries, with the largest samples to date being the China Kadoorie Biobank (a study with $\sim$500,000 genotyped individuals currently), Biobank Japan ($\sim$200,000 genotyped individuals currently), and the Taiwan Biobank ($\sim$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 (see, for example, 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).

4.7. *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 3.5 for formal definitions and discussion). These estimators rely on different assumptions and target different estimands (for example, narrow vs. broad 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 (for example, 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 reared apart have identical genetic factors but uncorrelated residuals, the correlation of their phenotypes is an estimator of broad 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 (for example, 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, for example, 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 large samples required for adequate power, these methods have rarely been applied using within-family identification (exceptions include Visscher et al., 2006, and Young et al., 2018). Instead, for identification these estimators rely on controlling for genetic principal components and are therefore sensitive to the confounds discussed in Sections 4.2 and 4.3.

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 (for example, 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 a SNP-level LD score, which measures how strongly that SNP is correlated with others 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.

Both genomic-relatedness methods and GWAS-based methods can have credible identification when applied to family samples (for example, applying a GWAS-based method to family-based GWAS summary statistics). This is due to the same quasi-experimental logic (and with the same caveats) as in Sections 4.4 and 4.5, 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). We expect that these study designs will become more common as larger family samples become available.

Even though largely applied to date with imperfect controls, 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).

5. *Polygenic Indexes*

Most applications using genetic data in the social sciences use a polygenic index (PGI), defined as a standardized, weighted sum of the genotypes of a set of observed genetic variants (typically SNPs):

$$g_{\mathbf{w}} := \frac{\tilde{\mathbf{x}}\mathbf{w}}{\mathrm{std}\left(\tilde{\mathbf{x}}\mathbf{w}\right)},$$

where $\tilde{\mathbf{x}}$ is the vector of observed genotypes, $\mathbf{w}$ is a vector of weights (the "PGI weights"), and $\mathrm{std}\left(\cdot\right)$ is the standard deviation of its argument across a population of individuals. 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; Alemu et al., 2025*b*)

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, if constructed and analyzed appropriately, a PGI can serve as an empirical proxy for the combined causal effects of observed SNPs—a partial measure of the "genetic propensity" for a phenotype.

To understand the sense in which a PGI can be a partial measure of "genetic propensity," as well as the "if constructed and analyzed appropriately" caveats, consider a typical social-science study that uses a PGI: researchers estimate the association between the PGI for educational attainment and some other individual-level outcome, such as mid-life income. Such a regression might be motivated by a model of the labor market returns to human capital, where the PGI for educational attainment—because it is fixed at conception and unaffected by human capital investments—is used as a measure of "initial ability." We discussed in Section 3.6 some conceptual caveats regarding genetic effects that extend to PGIs. One is that the PGI for educational attainment captures not only initial ability but also expected future investments. Another is that the PGI for educational attainment picks up genetic effects on other phenotypes that have a non-zero genetic correlation with educational attainment; examples include smoking and extraversion (Tan et al., 2024). In this section, we focus on additional considerations that are specific to a PGI.

Off the bat, because the PGI is a linear function of *observed* (measured and imputed) SNPs, at best the PGI can proxy for the additive SNP factor for educational attainment (as defined in Section 3.3)—not the genetic factor or additive genetic factor (as defined in Sections 3.1 and 3.2, respectively), unless strong additional assumptions are made (discussed in Section 5.5).[17] To be precise, the PGI would equal the standardized additive SNP factor for educational attainment if $\mathbf{w} = \tilde{\boldsymbol{\beta}}$, where $\tilde{\boldsymbol{\beta}}$ is the least-squares projection of causal genetic effects on educational attainment onto the observed SNPs.

In practice, researchers cannot set $\mathbf{w}$ equal to $\tilde{\boldsymbol{\beta}}$ because $\tilde{\boldsymbol{\beta}}$ is estimated, not known. Researchers instead proceed by setting $\mathbf{w}$ equal to an estimate of $\tilde{\boldsymbol{\beta}}$, which leads to two separate issues. First, instead of an unbiased estimate of $\tilde{\boldsymbol{\beta}}$, 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 estimates of $\tilde{\boldsymbol{\beta}}$. The extent to which the PGI can proxy for the additive SNP factor depends on how well $\mathbf{z}$ eliminates confounds. Second, the estimate is noisy. This noise generates approximately classical measurement error in the PGI, as we explain in Section 5.1. In Sections 5.1-5.3, we discuss the choice of weights and its consequences for PGI predictive power.

Once the researcher runs the regression of mid-life income on the PGI for educational attainment, a third issue arises: under what circumstances and in what sense does the coefficient on the PGI have a causal interpretation? This issue is orthogonal to the first two: as long as the regression is conducted within-family—for

[17] The discussion and analysis in this section apply also to PGI weights chosen such that the PGI approximates some quantity other than the additive SNP factor that can be expressed as a linear function of observed SNPs. For example, the genetic principal components estimated from a sample (discussed in Section 4.3) are PGIs that aim to approximate the population's (true) genetic principal components.

example, controlling for parental PGIs—it turns out that the coefficient can be interpreted as a weighted sum of causal effects of genetic variants, although the value of $\mathbf{w}$ determines *which* weighted sum. In Sections 5.4-5.6, we turn to this issue and other interpretational questions that arise when PGIs are analyzed in applications.

5.1. *PGI Weights and Interpretation of the PGI*

PGI weights $\mathbf{w}$ are generally constructed using the SNP coefficients estimated from a GWAS of some phenotype y. In Section 4, we motivated GWAS by considering a regression that is conceptually useful but not estimable. Here, to motivate the estimation and interpretation of PGI weights, we consider a similar regression; it differs from Equation (4) in that we make explicit that the regressors are the *observed* (measured or imputed) SNPs, $\tilde{\mathbf{X}}$, rather than all genetic variants, $\mathbf{X}$, and the regression includes the controls $\mathbf{Z}$:

$$\left\{\breve{\boldsymbol{\beta}}, \breve{\boldsymbol{\alpha}}\right\} \coloneqq \underset{\mathbf{b},\mathbf{a}}{\operatorname{argmin}} \left\{\mathbb{E}\left[\left(y - \tilde{\mathbf{X}}\mathbf{b} - \mathbf{Z}\mathbf{a}\right)^2\right]\right\}, \tag{7}$$

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 the coefficient on $\tilde{\mathbf{X}}$ as the *optimal predictor weights*, denoted $\breve{\boldsymbol{\beta}}$, to make clear that it may differ from the additive SNP coefficient $\tilde{\boldsymbol{\beta}}$ if the controls $\mathbf{z}$ are imperfect. We refer to $\breve{g} \coloneqq \tilde{\mathbf{x}}\breve{\boldsymbol{\beta}}$ as the *optimal predictor* (of the phenotype y). 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 of y is equal to the additive SNP factor for y; otherwise, they are generally not equal. Parallel to the notion of SNP heritability, we define $\breve{h}^2 \coloneqq \operatorname{Var}(\breve{g}) / \operatorname{Var}(y)$ as the *optimal predictive power* for y: 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 (see Section 4.7) on GWAS summary statistics that use $\mathbf{z}$ as the set of controls.

The sample analog of Equation (7) cannot be estimated for the same short-rank reasons discussed in Section 4. In practice, most of the commonly used estimators for $\breve{\boldsymbol{\beta}}$ are Bayesian and 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

[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 observed SNPs because all the observed SNPs could add information to the PGI. In addition

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, $\mathbf{\Sigma}$ (see, for example, 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 (Ni et al., 2021) but do not matter for the purposes of discussion here.

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

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

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 $\breve{g}$ (of y), with approximately classical measurement error. This result is central to the conceptual appeal of the PGI, and because the measurement error is approximately classical, there are relatively straightforward methods of correcting for the measurement error in applications (as discussed in Section 5.5).

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). To begin, we suppose that the GWAS, LD-estimation, and prediction samples are all drawn from the same population. We relax this assumption in Section 5.3 below.

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 $\breve{g} = \tilde{\mathbf{x}}\breve{\boldsymbol{\beta}}$:

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

where $e := \tilde{\mathbf{x}}\mathbf{u}$ 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 (7) above (which is not feasible because of the short-rank problem described in Section 4), 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 $\operatorname{Cov}(\breve{g}, e) = 0$ that $\operatorname{Var}(\breve{g} + e) = \operatorname{Var}(\breve{g}) + \operatorname{Var}(e)$. For a standard Bayesian approach to constructing a PGI discussed above, these properties do

to Bayesian approaches and pruning-and-thresholding, machine-learning approaches also exist (see, for example, Widen et al., 2021; Zhao et al., 2021).

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) derive formulas for these approximations and calculate 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 approximately 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.

5.2. *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 (for example, 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 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 now, we assume the two populations have a common LD matrix (for example, 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 B:

$$R^2 = \left(\breve{h}^2_{pred} r^2_{\mathbf{x}\boldsymbol{\beta}}\right) \left(\frac{\breve{h}^2_{GWAS}}{\breve{h}^2_{GWAS} + M/N}\right), \tag{8}$$

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" similar to but distinct from those defined in

Section 3.5 (for details, see Appendix B); M is a constant (discussed below); and N is the GWAS sample size underlying the PGI weights.

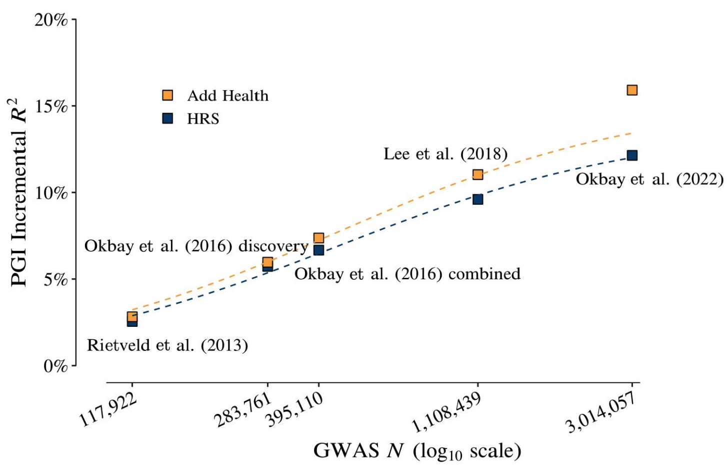
The first term in Equation (8)—the optimal predictive power for y_{pred} in the prediction population, $\breve{h}^2_{pred}$, multiplied by $r^2_{\mathbf{x}\boldsymbol{\beta}}$—is the predictive power that would be achieved if the PGI weights were estimated from an infinite GWAS sample ($N \to \infty$). 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, how they are measured (for example, 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 (8)—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 (8) 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 (8) 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 (to mitigate population-stratification bias; see Sections 4.2 and 4.3) 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 (8) is derived for a univariate regression, rather than for the incremental

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).

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 M in Equation (8) to the first four of the five points in the figure separately for each dataset. The functional form implied by Equation (8) provides a good fit to all five points in both datasets, except for the predictive power in Add Health from the most recent GWAS, which is larger than expected.

As another empirical illustration of Equation (8), 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 (8), for each PGI separately, Mostafavi et al. (2020) find that the incremental R^2 is larger when the GWAS 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. (Income and parental education are included because they are known to be among the strongest predictors of educational attainment, but note that outside a prediction context, they could be problematic controls if they pick up some channels through which genetic effects operate.) Second, in some settings, the incremental R^2 may not be the relevant measure of predictive power. For example, following Rietveld et al. (2013), we show in Appendix D that when a PGI is added as a control variable to increase precision in a randomized evaluation of an environmental intervention, the efficiency gains 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.

5.3. *PGI Portability Across Populations*

Many potential social-science applications involve data on African Americans, Hispanic, and other groups with genetic ancestries that typically differ from those 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 B, we generalize Equation (8) 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 for y_{GWAS} from the GWAS sample are no longer optimal for y_{GWAS} 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.

The sign of this third effect is indeterminate, but in most applications its expected magnitude is small.

In practice, the GWASs underlying PGI weights are typically conducted in samples of individuals of European genetic ancestries (see Section 4.6)—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) use 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 larger samples with genotype and sequencing data 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 differ in genetic ancestry, 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 unobserved 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) provide a striking empirical example: they compare the distributions of height PGIs for several different populations with different genetic ancestries. They find 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.

5.4. *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 subtleties arise because 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, hypothetical experiments that involve modifying genotypes at conception define average treatment effects that are not estimable (fundamentally because, in the data, only individuals who have at least one heterozygous parent contribute identifying variation at any given genotype and thus what is identified is a local average treatment effect, as discussed in Section 4.6). To illustrate the kinds of challenges that arise, consider another candidate hypothetical experiment: 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. Thus, the association between the PGIs and the potential outcomes has a causal interpretation. Here, however, the problem is that it is an average treatment effect across embryos *conditional on the parents*, and it is not clear how that relates to a quantity that could be estimated.

Another approach would be to build up from the definition of the causal effect of a genetic variant to a definition of the "causal effect of a PGI" by appropriately adding up the causal effects of relevant variants.That approach runs into two conceptual challenges. 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 unobserved genetic variants that are correlated with observed SNPs. (For this same reason, in Section 3.3, 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, two individuals whose PGIs changed by the same amount might have different hypothetical experiments that define the effect.[19]

Rather than taking any of the above approaches, like Veller, Przeworski and Coop (2024), we work backwards from a definition of 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 the 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 4.4, an individual's genotype vector, $\mathbf{x}$, can be decomposed into the mean parental genotype vector, $\mathbf{x}_p := \frac{\mathbf{x}_f+\mathbf{x}_m}{2}$, and a random deviation, $\mathbf{x}_r := \mathbf{x} - \frac{\mathbf{x}_f+\mathbf{x}_m}{2}$. Consider the population regression of some outcome variable y on $\mathbf{x}$ and $\mathbf{x}_p$,

$$y = \mathbf{x}\boldsymbol{\beta} + \mathbf{x}_p\mathbf{b}_p + \xi. \tag{9}$$

As in the single-SNP case discussed in Section 4.6, the coefficient on the child's genotype vector, $\boldsymbol{\beta}$, is the vector of causal genetic effects (or more precisely, local average treatment effects because it is identified from children with at least heterozygous parent), whereas 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

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

[19] Which genotypes were changed would not matter if two conditions are both satisfied: (i) the additive model is true (that is, 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 for educational attainment by one unit. If one of the SNPs affects income and the other does not, the two hypothetical experiments have different effects.

where (unlike in Section 5.1 above) we now express the PGI as a weighted sum of the genotypes of *all* genetic variants, with $w_j = 0$ for every unobserved 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, $p_{\mathbf{w}} \coloneqq \mathbf{x}_f\mathbf{w} + \mathbf{x}_m\mathbf{w}$:

$$y = \alpha_g g_{\mathbf{w}} + \alpha_p p_{\mathbf{w}} + u \tag{10}$$

(where we define $p_{\mathbf{w}}$ as a sum rather than an average because it makes the expressions for α_p and α_g symmetric). In Appendix C, we derive the relationship between the coefficients from the regressions in equations (9) and (10). 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:

$$\alpha_g = \frac{\mathbf{w}'\boldsymbol{\Sigma}\boldsymbol{\beta}}{\mathbf{w}'\boldsymbol{\Sigma}\mathbf{w}} \tag{11}$$

$$\alpha_p = \frac{\mathbf{w}'\boldsymbol{\Sigma}\mathbf{b}_p}{\mathbf{w}'\boldsymbol{\Sigma}\mathbf{w}}, \tag{12}$$

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

Equation (11) 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 (11) 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 $\mathbf{x}\boldsymbol{\beta}$—which is related to the additive genetic factor for the outcome y but using the local-average-treatment-effect coefficients in place of the average-treatment-effect coefficients—on the PGI, estimated using only the random component of the genotype vector. For example, if y is income, the PGI is for educational attainment, and the local average treatment effects for income are equal to the average treatment effects, then α_g measures the slope of the relationship between the additive genetic factor for income and the PGI for educational attainment, given the distribution of exogenous variation in the PGI in the population. This is most straightforward to see if we ignore the parental PGI in regression Equation (10) and treat the genotype vector as exogenous.

Then, regressing the additive genetic factor, $\mathbf{x}\boldsymbol{\beta}$, on the PGI gives coefficient

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

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 (that is, more common SNPs) and SNPs that are more strongly correlated with other genetic variants (including unobserved variants) because these SNPs are responsible for more of the variance in the additive genetic factor. This formulation makes clear that α_g will be non-zero if the causal genetic effects on income are correlated with the PGI weights for educational attainment, even if some or all of the causal effects on income do not operate through educational attainment.

Our analysis in Appendix C 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 (12) 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.[20]

Just as $\boldsymbol{\beta}$ is identified if regression Equation (9) includes the father's and mother's genotypes separately rather than the mean parental genotype, α_g is identified if regression Equation (10) 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 (10), 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. When including the father's and mother's PGIs separately and more generally when 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 an-

[20]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.

alyze sibling samples and control for family fixed effects. As in the case of estimating the effects of genotypes discussed in Section 4.5, 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: $g_{\mathbf{w}} - \left(g_{\mathbf{w}} + g_{\mathbf{w},sib}\right)/2 = \left(g_{\mathbf{w}} - g_{\mathbf{w},sib}\right)/2 = \left(\mathbf{x} - \mathbf{x}_{sib}\right)\mathbf{w}/2$, where the subscript "sib" denotes an individual's sibling. Variation in $\left(g_{\mathbf{w}} - g_{\mathbf{w},sib}\right)/2$ is random, but in a regression of y on $g_{\mathbf{w}}$ with family fixed effects, the coefficient is

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

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 lower PGI.

Analogous to the discussion in Section 4.4, when sibling genotypes are observed but parental genotypes are not, Young et al. (2022) show 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 substantially improves precision over the family-fixed-effects specification while preserving credible identification of Equation (11).

To date, most applications involving PGIs have estimated a regression like Equation (10) but with controls $\mathbf{z}$, such as principal components of the genetic data, that may not fully capture all confounds instead of controlling for the sum of parental PGIs. Such a regression can be understood as estimating the parameter in Equation (11), but with omitted-variables bias due to uncorrected-for gene-environment correlation, as well as bias due to assortative mating. In some cases, these biases can be substantial. For example, Okbay et al. (2022) compare, for 23 phenotypes, the coefficient from a regression of the phenotype on a PGI for educational attainment controlling for principal components with the coefficient from the same regression but instead controlling for parental PGIs. On average across the 23 phenotypes, the coefficient controlling for parental PGIs is 59% (SE = 1%) as large as the coefficient controlling for principal components. For the 13 out of 23 phenotypes—including educational attainment itself, cognitive performance, and personal income—where the two estimates are statistically distinguishable (after Bonferonni correction for 23 tests), the coefficient controlling for parental PGIs is closer to zero. This evidence suggests that, in general, the omitted-variables and assortative-mating biases in the estimates that control for principal components go in the direction of exaggerating the estimated effect of the PGI.

5.5. *Correcting for PGI Measurement Error in Applications*

As we showed in Section 5.1, the PGI for some phenotype can be interpreted as a standardized, noisy measure of the optimal predictor of that phenotype—the coefficient $\breve{\boldsymbol{\beta}}$ in Equation (7)—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 (as discussed, for example, by Gillen, Snowberg and Yariv, 2019). Moreover, since PGIs vary by GWAS source and are constructed using different methods and different SNPs, 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) propose an instrumental-variables approach. Two independent subsamples of the GWAS are used to construct two sets of PGI weights. From these weights, two independent PGIs are constructed in the prediction sample. Each PGI is then 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 methods discussed in Section 4.7, 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) provide a detailed analysis (but focus on the case of a regression of some phenotype on the PGI for that same phenotype, which is rarely the case of interest in applications).[21]

As discussed at the beginning of Section 5, researchers typically wish to interpret a PGI for a phenotype as an unbiased estimate of the additive SNP factor for that phenotype—but as also discussed, this interpretation is valid only if the GWAS controls are sufficient for a credible causal interpretation. If we grant this premise, a natural question arises: can we further adjust our estimates to

[21] 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.

approximate what the results would look like if the additive genetic factor had been analyzed instead? This would generally align better with theoretical models that motivate applications. Such a correction is indeed feasible but requires two additional assumptions: that narrow heritability is known and that the portion of the additive genetic factor that is not captured by the additive SNP factor has the same effect on the outcome being studied as the effect of the additive SNP factor. The latter assumption is required because a PGI can only contain information about the additive SNP factor, so an adjustment that seeks to recover the effect of the whole additive genetic factor must extrapolate to the effects of unobserved variants. In practice, neither assumption is likely to hold exactly. For example, the relative-effect-magnitude assumption is likely to fail because the unobserved variants that have low correlation with observed SNPs are primarily rare variants, whose effect sizes can differ appreciably from those of common SNPs for a variety of reasons. While applying the correction might still be of interest, it is essential to be transparent about the underlying assumptions.

5.6. *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 3.4 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. These mechanisms may 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 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.

6. *Applications*

In this section, we critically discuss applications in economics that make use of genetic data. While economists began working with genetic data earlier (see, for example, Fletcher and Lehrer, 2009), most of the recent work has analyzed PGIs. Influential contributions include Papageorge and Thom (2020) and Barth, Papageorge and Thom (2020). Papageorge and Thom (2020) address the long-standing 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 introduce the idea of using instead a PGI for educational attainment, which is fixed at conception and therefore predetermined with respect to any human capital investments. The paper finds that the association between the PGI and educational outcomes is substantially stronger for individuals who grew up in households with higher socioeconomic status. The papers also finds that the PGI predicts labor earnings even after controlling for completed education, with a larger relationship in more recent decades.

The contribution by 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 (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 one–standard-deviation increase in the PGI for educational attainment predicts 23% higher household wealth, controlling for education, income, and business ownership, suggesting that the PGI captures differences in financial decision-making. Consistent with this interpretation, a higher PGI is associated with stock market participation, financial literacy, and planning behavior, but it does *not* predict wealth among households with defined-benefit pensions, who have little discretion over financial choices. As papers we discuss below will illustrate, some of the subsequent literature has taken up the same research topics as these early papers and built 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 applications 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.

6.1. *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 (for example, 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 (including Barcellos et al. 2018; 2025, and Schmitz and Conley, 2017). For instance, Barcellos, Carvalho and Turley (2018) (their Appendix B) use 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, as in studies of the association between parental behaviors and children's outcomes (see, for example, 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).[22]

Rietveld et al. (2013) calculate the gains in effective sample size that could be obtained by controlling for PGIs in a simple RCT with two conditions (see Appendix D for details). For example, if the set of baseline controls, absent the PGI, explains 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 (for example, 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.

6.2. *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 that 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 6.6). 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

[22] 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).

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 randomly or quasi-randomly assigned, 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 5.5 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 (observed 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 for example, 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 develop a powerful framework for conceptualizing and estimating gene-by-environment interactions without using PGIs. Building on the assumptions and approach of LD Score regression (see Section 4.7), 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) discussed 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.

6.3. *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 (8). 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 (for example, 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 (for example, 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.

6.4. *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 (for example, 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) make this observation and exploit 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, for example, 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.

6.5. *Interpersonal Genetic Effects*

As discussed in Section 3.7, 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) examine the association of school friends' PGIs for height, BMI, and educational attainment with one's own phenotype values and find 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 (Armstrong-Carter et al., 2020). Domingue and Fletcher (2020) find that the association with children's educational attainment is stronger for biological children than 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: 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. We discuss why the coefficients on parents' PGIs cannot be interpreted causally in Section 5.4.

The best-identified work to date we are aware of is Young et al. (2022), who estimate sibling genetic effects by regressing the individual's phenotype on their own PGI, their sibling's PGI, and the (observed 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 and friends' parental PGIs, respectively. 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.

6.6. *Mendelian Randomization*

There is a vast epidemiological literature that uses genetic variants as instrumental variables to make causal inferences about the effects of interventions, called "exposures" in this literature, 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); a recent example is Widding-Havneraas et al. (2026). For skeptical reactions, see, for example, Conley (2009), Cawley, Han and Norton (2011), Beauchamp et al. (2011), Benjamin et al. (2012), and McMartin and Conley (2020). For a recent overview, see Sanderson et al. (2026). 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 criteria 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 by which the variant could be correlated with the outcome—for example, those arising from population stratification or from causal effects of the variant on more than one phenotype (often called *horizontal pleiotropy*[23]). It is virtually never possible to rule out all alternative pathways because the variant may be in LD with many other variants that could affect the outcome. Typically, the best-case scenario is that the known effect of the variant on the exposure is direct and large enough that it is plausibly much larger than effects through alternative pathways. Third, results from placebo tests, sometimes called "negative controls," conducted in subpopulations where the exposure is absent can provide a check. 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 criteria 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), who investigate 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

[23] 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 neither phenotype causes the other, for example, because the genetic variant affects the phenotypes through distinct biological mechanisms. "Vertical pleiotropy" refers to when a genetic variant affects one phenotype, A, that in turn affects another phenotype, B. While horizontal pleiotropy may lead to violation of the exclusion restriction, vertical pleiotropy implies that the exclusion restriction holds for studying the effect of A on B.

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 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 (see, for example, 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, mainly because, as noted above, researchers can virtually never rule out LD between the genetic variant used as the instrument and other genetic variants that may affect the outcome. Nonetheless, MR studies can be valuable in settings where, relative to the biases from other feasible study designs, the biases from violations of the MR assumptions are likely to be small.

6.7. *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 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 4.4) 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

$$I_i = \beta_0 + \beta_1 \left(\mathrm{PGI}_i - \mathrm{PGI}_{sib}\right) + \beta_2 \mathrm{PGI}_i + \beta_3 \mathrm{PGI}_{par} + \mathrm{Controls} + u, \tag{13}$$

where I_i is the index of parental investment in the focal child, PGI_i and PGI_{sib} are the focal child's and sibling's respective PGIs, PGI_{par} is the (imputed) parental mean PGI, and u is an error term (and standard errors are clustered by family). Because PGI_i and $(\mathrm{PGI}_i - \mathrm{PGI}_{sib})$ 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. As discussed in Section 5.5, estimating regression Equation (13) 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 (for example, 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 (for example, 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 (for example, 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 4.6). 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 3.3 and 5.4). 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 5.5), doing so is well within the spirit of structural modeling.

7. *Future Directions and Concluding Remarks*

Over the last ten years, with the advent of genome-wide association studies (GWASs) for social and behavioral phenotypes—which have enabled researchers to identify an increasing fraction of the single-nucleotide polymorphisms (SNPs) associated with these phenotypes—and especially family-based GWASs—which enable researchers to identify *causal* genetic effects—social-science genomics has come of age. Polygenic indexes (PGIs), which aggregate the information uncovered by GWAS into variables that can be readily incorporated into standard regressions, are beginning to be used in social-science applications, as illustrated by the examples in Section 6. 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, such as those showcased in Section 6.7, economists may contribute to geneticists' understanding of the mechanisms through which PGIs matter.

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-by-environment interactions.

Second, as larger genotyped family samples become available, researchers will be able to more precisely estimate—with credible causal identification—not only effects of an individual's genotype on their own phenotypes but also interpersonal genetic effects, such as effects of an individual's genotype on their family members' phenotypes (see Sections 3.7 and 6.5). 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. This shift will improve applications because the latter is typically closer to the theoretical construct of interest (see Sections 5.1 and 5.5).

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, thereby substantially expanding the range of possible applications.

Fourth, genotyping will become denser, implying that the observed 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, for example, on intellectual disabilities (for example, Chen et al., 2023), that evade detection in association studies limited to (relatively common) SNPs. 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 (see Sections 3.2 and 5.5), which more closely approximates the theoretical construct of interest in many social-science applications than do current PGIs.

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 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 (for example, 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.

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, 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.

Björklund, Anders, Markus Jäntti, and Gary Solon. 2005. "Influences of Nature and Nurture on Earnings Variation." In *Unequal Chances: Family Background and Economic Success.* , ed. Samuel Bowles, Herbert Gintis and Melissa Osborne Groves, 145–165. Princeton University Press.

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.

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.

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.

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.

Ni, Guiyan, Jian Zeng, Joana A. Revez, Ying Wang, Zhili Zheng, et al. 2021. "A Comparison of Ten Polygenic Score Methods for Psychiatric Disorders Applied Across Multiple Cohorts." *Biological Psychiatry*, 90(9): 611–620.

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., Dalton Conley, Nicholas Eriksson, Tõnu Esko, Sarah E. Medland, et al. 2014. "Replicability and Robustness of GWAS for Behavioral Traits." *Psychological Science*, 25(11): 1975–1986.

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.

Sanderson, Eleanor, Michael G. Levin, Venexia Walker, Shuai Yuan, Isabella Badini, Julia Dolce, Karina J. Mahida, Ju-Woo Nho, Jean-Baptiste Pingault, Scott M. Damrauer, Gibran Hemani, and Neil M. Davies. 2026. “Challenges and future directions for Mendelian randomization.” *Nature genetics.*

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.

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*.

Widding-Havneraas, Tarjei, Perline A. Demange, Henrik Daae Zachrisson, Nicolai Borgen, Eivind Ystrøm, and Felix Elwert. 2026. "Estimating returns to education using the genetic lottery." *Proceedings of the National Academy of Sciences of the United States of America*, 123: e2537049123.

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.

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 Disequilibrium 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.

Appendix A. *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} (that is, the weighted sum of causal effects of the SNP together with the causal effects of genetic variants correlated with the SNP, following the standard omitted-variables bias formula) 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

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

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

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

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

Next, we evaluate the reduced-form term and obtain

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

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.

Appendix B. *Derivations of Formulae for PGI Predictive Power*

Here, we derive analytic formulae for the predictive power of a PGI, beginning with Equation (8) in the main text (de Vlaming et al., 2017). Letting $\breve{g}_{\text{pred}}$ denote the optimal predictor in the prediction population:

$$R^2 = \frac{\operatorname{Cov}\left(y_{\text{pred}}, \hat{g}\right)^2}{\operatorname{Var}\left(y_{\text{pred}}\right)\,\operatorname{Var}\left(\hat{g}\right)} = \frac{\operatorname{Cov}\left(y_{\text{pred}}, \dfrac{\breve{g}+e}{\operatorname{sd}(\breve{g}+e)}\right)^2}{\operatorname{Var}\left(y_{\text{pred}}\right)} = \frac{\operatorname{Cov}\left(y_{\text{pred}}, \breve{g}\right)^2}{\operatorname{Var}\left(y_{\text{pred}}\right)\left[\operatorname{Var}(\breve{g}) + \operatorname{Var}(e)\right]}$$

$$= \left(\frac{\operatorname{Cov}\left(y_{\text{pred}}, \breve{g}\right)^2}{\operatorname{Var}\left(y_{\text{pred}}\right)\,\operatorname{Var}(\breve{g})}\right)\left(\frac{\operatorname{Var}(\breve{g})}{\operatorname{Var}(\breve{g}) + \operatorname{Var}(e)}\right)$$

$$= \left(\frac{\operatorname{Var}\left(\breve{g}_{\text{pred}}\right)}{\operatorname{Var}\left(y_{\text{pred}}\right)} \cdot \frac{\operatorname{Cov}\left(y_{\text{pred}}, \breve{g}\right)^2}{\operatorname{Var}\left(\breve{g}_{\text{pred}}\right)\,\operatorname{Var}(\breve{g})}\right)\left(\frac{\breve{h}^2_{\text{GWAS}}}{\breve{h}^2_{\text{GWAS}} + \operatorname{Var}(e)/\operatorname{Var}\left(y_{\text{GWAS}}\right)}\right)$$

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

where M is a constant and N is the GWAS sample size underlying the PGI weights; the derivation relies on the approximation $\operatorname{Cov}(y, e) = 0$ discussed in main text Section 5.1. The second-to-last equality follows from the definition of the optimal predictor, which implies $\operatorname{Cov}(y_{\text{pred}}, \breve{g}) = \operatorname{Cov}(\breve{g}_{\text{pred}} + \varepsilon_{\text{pred}}, \breve{g}) = \operatorname{Cov}(\breve{g}_{\text{pred}}, \breve{g})$, where $\varepsilon_{\text{pred}}$ is the error in predicting y_{pred} in the prediction population using $\breve{g}_{\text{pred}}$. The last equality follows because $\operatorname{Var}(e)$ 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 relax this assumption but do 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) relax the assumption without making parametric assumptions but do not formally define and interpret all the parameters.[1]

We begin by establishing some notation. First, let

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

Here, y_{pred} is the phenotype in the prediction population, $\tilde{\mathbf{x}}_{\text{pred}}$ is the vector of observed SNP genotypes, $\breve{\boldsymbol{\beta}}_{\text{pred}}$ is the vector of optimal predictor weights for y_{pred},

[1] 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.

and $\tilde{\varepsilon}_{\text{pred}}$ is a residual that is uncorrelated with the genotypes. Next, let

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

denote a PGI constructed in the prediction population using estimates of PGI weights from the GWAS population, $\hat{\boldsymbol{\beta}}_{\text{GWAS}}$, and let $\mathbf{u}_{\text{GWAS}}$ denote the estimation error from such a projection in a finite sample. Finally, let $\boldsymbol{\Sigma}_{\text{pred}} \coloneqq \text{Var}\left(\tilde{\mathbf{x}}_{\text{pred}}\right)$ and $\boldsymbol{\Sigma}_{\text{GWAS}} \coloneqq \text{Var}\left(\tilde{\mathbf{x}}_{\text{GWAS}}\right)$ 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:

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

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

$$R^2 = \breve{h}^2_{\text{pred}} r_g^2 \frac{\text{Var}\left(\tilde{\mathbf{x}}_{\text{pred}}\breve{\boldsymbol{\beta}}_{\text{pred}}\right)}{\text{Var}\left(\tilde{\mathbf{x}}_{\text{pred}}\hat{\boldsymbol{\beta}}_{\text{GWAS}}\right)}. \tag{B1}$$

For some intuition on how to interpret the parameter r_g^2, consider first the special case when $\boldsymbol{\Sigma}_{\text{pred}} = \boldsymbol{\Sigma}_{\text{GWAS}}$, $\breve{\boldsymbol{\beta}}_{\text{GWAS}} = \tilde{\boldsymbol{\beta}}_{\text{GWAS}}$, and $\breve{\boldsymbol{\beta}}_{\text{pred}} = \tilde{\boldsymbol{\beta}}_{\text{pred}}$. Then r_g^2 is the squared correlation between two additive SNP factors, one based on the GWAS weights ($\tilde{\boldsymbol{\beta}}_{\text{GWAS}}$) and one based on the prediction sample weights ($\tilde{\boldsymbol{\beta}}_{\text{pred}}$), so r_g is an instance of the genetic correlation parameter $r_{\boldsymbol{x\beta}}$. In the more general case when $\boldsymbol{\Sigma}_{\text{pred}} \neq \boldsymbol{\Sigma}_{\text{GWAS}}$, $\breve{\boldsymbol{\beta}}_{\text{GWAS}} \neq \tilde{\boldsymbol{\beta}}_{\text{GWAS}}$, and $\breve{\boldsymbol{\beta}}_{\text{pred}} \neq \tilde{\boldsymbol{\beta}}_{\text{pred}}$, 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 $\mathrm{Var}\big(\tilde{\mathbf{x}}_{\mathrm{pred}}\hat{\boldsymbol{\beta}}_{\mathrm{GWAS}}\big)=\mathrm{Var}\big(\tilde{\mathbf{x}}_{\mathrm{pred}}\hat{\boldsymbol{\beta}}_{\mathrm{GWAS}}\big)+\mathrm{Var}\big(\tilde{\mathbf{x}}_{\mathrm{pred}}\mathbf{u}_{\mathrm{GWAS}}\big)$, Equation (B1) can be rewritten as:

$$
\begin{aligned}
R^2 &= \breve{h}^2_{\mathrm{pred}} r_g^2 \frac{\mathrm{Var}\left(\tilde{\mathbf{x}}_{\mathrm{pred}}\breve{\boldsymbol{\beta}}_{\mathrm{pred}}\right)}{\mathrm{Var}\big(\tilde{\mathbf{x}}_{\mathrm{pred}}\hat{\boldsymbol{\beta}}_{\mathrm{GWAS}}\big)+\mathrm{Var}\big(\tilde{\mathbf{x}}_{\mathrm{pred}}\mathbf{u}_{\mathrm{GWAS}}\big)} \\
&= \breve{h}^2_{\mathrm{pred}} r_g^2 \left(\frac{\frac{\mathrm{Var}\left(\tilde{\mathbf{x}}_{\mathrm{pred}}\hat{\boldsymbol{\beta}}_{\mathrm{GWAS}}\right)}{\mathrm{Var}(y_{\mathrm{GWAS}})}}{\frac{\mathrm{Var}\left(\tilde{\mathbf{x}}_{\mathrm{pred}}\hat{\boldsymbol{\beta}}_{\mathrm{GWAS}}\right)}{\mathrm{Var}(y_{\mathrm{GWAS}})}+\frac{\mathrm{Var}\left(\tilde{\mathbf{x}}_{\mathrm{pred}}\mathbf{u}_{\mathrm{GWAS}}\right)}{\mathrm{Var}(y_{\mathrm{GWAS}})}} \right). \qquad \text{(B2)}
\end{aligned}
$$

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

$$
\begin{aligned}
\frac{\mathrm{Var}\left(\tilde{\mathbf{x}}_{\mathrm{pred}}\breve{\boldsymbol{\beta}}_{\mathrm{GWAS}}\right)}{\mathrm{Var}\left(y_{\mathrm{GWAS}}\right)} &= \frac{\breve{\boldsymbol{\beta}}'_{\mathrm{GWAS}}\boldsymbol{\Sigma}_{\mathrm{pred}}\breve{\boldsymbol{\beta}}_{\mathrm{GWAS}}}{\mathrm{Var}\left(y_{\mathrm{GWAS}}\right)} \\
&= \frac{\breve{\boldsymbol{\beta}}'_{\mathrm{GWAS}}\left(\boldsymbol{\Sigma}_{\mathrm{pred}}-\boldsymbol{\Sigma}_{\mathrm{GWAS}}+\boldsymbol{\Sigma}_{\mathrm{GWAS}}\right)\breve{\boldsymbol{\beta}}_{\mathrm{GWAS}}}{\mathrm{Var}\left(y_{\mathrm{GWAS}}\right)} \\
&= \breve{h}^2_{\mathrm{GWAS}} + \frac{\breve{\boldsymbol{\beta}}'_{\mathrm{GWAS}}\left(\boldsymbol{\Sigma}_{\mathrm{pred}}-\boldsymbol{\Sigma}_{\mathrm{GWAS}}\right)\breve{\boldsymbol{\beta}}_{\mathrm{GWAS}}}{\mathrm{Var}\left(y_{\mathrm{GWAS}}\right)}.
\end{aligned}
$$

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

$$
\frac{\mathrm{Var}\left(\tilde{\mathbf{x}}_{\mathrm{pred}}\breve{\boldsymbol{\beta}}_{\mathrm{GWAS}}\right)}{\mathrm{Var}\left(y_{\mathrm{GWAS}}\right)} = \frac{\breve{\boldsymbol{\beta}}'_{\mathrm{GWAS}}\Delta_{\boldsymbol{\Sigma}}\breve{\boldsymbol{\beta}}_{\mathrm{GWAS}}}{\mathrm{Var}\left(y_{\mathrm{GWAS}}\right)}. \qquad \text{(B3)}
$$

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

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

This approximation requires that the GWAS association of each SNP be small such that $\mathrm{Var}\left(y_{\mathrm{GWAS}}\right)$ 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, $\tilde{\mathbf{x}}_{\mathrm{pred}}$ and $\mathbf{u}_{\mathrm{GWAS}}$ are mean-zero and independent. The second term in the denominator can therefore be expressed as follows:

$$\text{(B4)} \qquad \frac{\mathrm{Var}(\tilde{\mathbf{x}}_{\mathrm{pred}}\mathbf{u}_{\mathrm{GWAS}})}{\mathrm{Var}(y_{\mathrm{GWAS}})} = \frac{\sum[\mathrm{Var}(\tilde{\mathbf{x}}_{\mathrm{pred}}) \circ \mathrm{Var}(\mathbf{u}_{\mathrm{GWAS}})]}{\mathrm{Var}(y_{\mathrm{GWAS}})}$$

$$\approx \frac{\sum\left[\mathrm{Var}(\tilde{\mathbf{x}}_{\mathrm{pred}}) \circ \frac{\mathrm{Var}(y_{\mathrm{GWAS}})}{N}\,\boldsymbol{\Sigma}_{\mathrm{GWAS}}^{-1}\right]}{\mathrm{Var}(y_{\mathrm{GWAS}})}$$

$$= \frac{1}{N}\sum\left(\boldsymbol{\Sigma}_{\mathrm{pred}} \circ \boldsymbol{\Sigma}_{\mathrm{GWAS}}^{-1}\right).$$

where $\circ$ denotes the element-wise multiplication operator and $\sum(\cdot)$ denotes the grand sum (that is, the sum over all the elements of the matrix). Substituting Equations (B3) and (B4) into (B2) and rearranging yields the following generalized formula for R^2:

$$\text{(B5)} \qquad \breve{h}^2_{\mathrm{pred}}\, r_g^2 \left(\frac{\breve{h}^2_{\mathrm{GWAS}} + \frac{(\breve{\boldsymbol{\beta}}_{\mathrm{GWAS}})'\,\Delta_{\boldsymbol{\Sigma}}\,\breve{\boldsymbol{\beta}}_{\mathrm{GWAS}}}{\mathrm{Var}(y_{\mathrm{GWAS}})}}{\breve{h}^2_{\mathrm{GWAS}} + \frac{(\breve{\boldsymbol{\beta}}_{\mathrm{GWAS}})'\,\Delta_{\boldsymbol{\Sigma}}\,\breve{\boldsymbol{\beta}}_{\mathrm{GWAS}}}{\mathrm{Var}(y_{\mathrm{GWAS}})} + \frac{1}{N}\sum\left(\boldsymbol{\Sigma}_{\mathrm{pred}} \circ \boldsymbol{\Sigma}_{\mathrm{GWAS}}^{-1}\right)} \right).$$

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 (B4) can be expressed as:

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

In what follows, we will treat $\sigma^2_{\mathrm{pred},j}$ and $\sigma^2_{\mathrm{GWAS},j}$ as identically distributed random variables[2] and examine two benchmark cases: one in which $\sigma^2_{\mathrm{pred},j}=\sigma^2_{\mathrm{GWAS},j}$ 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

[2]We believe this assumption is reasonable for PGIs constructed using the Bayesian methods we focus on in this paper which use all observed 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 in the "pruning-and-thresholding" approach, 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.

there are no natural selection forces that cause allele frequencies to be similar. Under first scenario, we obtain

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

consistent with the analytical results reported in Daetwyler, Villanueva and Woolliams (2008) and de Vlaming et al. (2017). To see this, note that if $\sigma^2_{\text{pred},j} = \sigma^2_{\text{GWAS},j}$, then $\Delta_{\mathbf{\Sigma}}$ is a null matrix and $r_g = r_{\boldsymbol{x\beta}}$. Therefore,

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

which is exactly Equation (8) in the main text. Under the second scenario, the expected value of the $\text{Var}\left(\tilde{\mathbf{x}}_{\text{pred}}\mathbf{u}_{\text{GWAS}}\right)/\text{Var}\left(y_{\text{GWAS}}\right)$ term is:

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

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 $\text{Var}\left(\tilde{\mathbf{x}}_{\text{pred}}\mathbf{u}_{\text{GWAS}}\right) / \text{Var}\left(y_{\text{GWAS}}\right)$.

Appendix C. *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 in 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.[3]

Let $\mathbf{x}$ denote a vector of genotypes for some individual, $\mathbf{x}_m$ denote the vector of genotypes for the individual's mother, and $\mathbf{x}_f$ denote the vector of genotypes of the individual'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:

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

[3] 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.

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

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

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

Using these results,

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

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, values of $1/2$ for any pair of genotypes corresponding to different chromosomes, and 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,

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

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

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

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

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

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 A). 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\mathbf{\Sigma} + (\mathbf{1}-\mathbf{P})\circ\mathbf{B} + \mathbf{A}\right]\mathbf{w} & 2\mathbf{w}'\left[(\mathbf{\Sigma}+\mathbf{B}) + 2\mathbf{A}\right]\mathbf{w} \\ & 4\mathbf{w}'\left[(\mathbf{\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\mathbf{\Sigma} + (\mathbf{1}-\mathbf{P})\circ\mathbf{B} + \mathbf{A}\right]\mathbf{w} & 2\mathbf{w}'\left[(\mathbf{\Sigma}+\mathbf{B}) + 2\mathbf{A}\right]\mathbf{w} \\ & 4\mathbf{w}'\left[(\mathbf{\Sigma}+\mathbf{B}) + 2\mathbf{A}\right]\mathbf{w} \end{bmatrix}^{-1} \\
&\times \begin{bmatrix} 2\mathbf{w}'\left[\mathbf{P}\circ\mathbf{\Sigma} + (\mathbf{1}-\mathbf{P})\circ\mathbf{B} + \mathbf{A}\right]\boldsymbol{\beta} + 2\mathbf{w}'(\mathbf{\Sigma}+\mathbf{B}+2\mathbf{A})\mathbf{b}_p \\ 2\mathbf{w}'\left[\mathbf{\Sigma}+\mathbf{B}+2\mathbf{A}\right]\boldsymbol{\beta} + 4\mathbf{w}'(\mathbf{\Sigma}+\mathbf{B}+2\mathbf{A})\mathbf{b}_p \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{w}'\left[\mathbf{P}\circ\mathbf{\Sigma} + (\mathbf{1}-\mathbf{P})\circ\mathbf{B} + \mathbf{A}\right]\mathbf{w} & \mathbf{w}'\left[(\mathbf{\Sigma}+\mathbf{B}) + 2\mathbf{A}\right]\mathbf{w} \\ & 2\mathbf{w}'\left[(\mathbf{\Sigma}+\mathbf{B}) + 2\mathbf{A}\right]\mathbf{w} \end{bmatrix}^{-1} \\
&\times \begin{bmatrix} \mathbf{w}'\left[\mathbf{P}\circ\mathbf{\Sigma} + (\mathbf{1}-\mathbf{P})\circ\mathbf{B} + \mathbf{A}\right]\boldsymbol{\beta} + \mathbf{w}'(\mathbf{\Sigma}+\mathbf{B}+2\mathbf{A})\mathbf{b}_p \\ \mathbf{w}'\left[\mathbf{\Sigma}+\mathbf{B}+2\mathbf{A}\right]\boldsymbol{\beta} + 2\mathbf{w}'(\mathbf{\Sigma}+\mathbf{B}+2\mathbf{A})\mathbf{b}_p \end{bmatrix} \\
&= \frac{1}{D}\begin{bmatrix} 2\mathbf{w}'\left[(\mathbf{\Sigma}+\mathbf{B}) + 2\mathbf{A}\right]\mathbf{w} & -\mathbf{w}'\left[(\mathbf{\Sigma}+\mathbf{B}) + 2\mathbf{A}\right]\mathbf{w} \\ & \mathbf{w}'\left[\mathbf{P}\circ\mathbf{\Sigma} + (1-\mathbf{P})\circ\mathbf{B} + \mathbf{A}\right]\mathbf{w} \end{bmatrix} \\
&\times \begin{bmatrix} \mathbf{w}'\left[\mathbf{P}\circ\mathbf{\Sigma} + (\mathbf{1}-\mathbf{P})\circ\mathbf{B} + \mathbf{A}\right]\boldsymbol{\beta} + \mathbf{w}'(\mathbf{\Sigma}+\mathbf{B}+2\mathbf{A})\mathbf{b}_p \\ \mathbf{w}'\left[\mathbf{\Sigma}+\mathbf{B}+2\mathbf{A}\right]\boldsymbol{\beta} + 2\mathbf{w}'(\mathbf{\Sigma}+\mathbf{B}+2\mathbf{A})\mathbf{b}_p \end{bmatrix}
\end{aligned}$$

where, after simplifying,

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

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

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

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 $(2\mathbf{P}-\mathbf{1})\circ(\mathbf{\Sigma}-\mathbf{B})$. This dispersion matrix is the variance-covariance matrix of the deviation of the child's genotypes from the mean parental genotype, $\mathbf{x}_r = \mathbf{x} - (1/2)\mathbf{x}_p$. This can be shown by taking the variance of $\mathbf{x}$, solving for $\operatorname{Var}(\mathbf{x}_r)$, and substituting the expressions for $\operatorname{Var}(\mathbf{x})$

and $\operatorname{Var}(\mathbf{x}_p)$ earlier in the section:

$$\begin{aligned}
\operatorname{Var}(\mathbf{x}) &= \operatorname{Var}\left(\frac{1}{2}\mathbf{x}_p\right) + \operatorname{Var}(\mathbf{x}_r) \\
\operatorname{Var}(\mathbf{x}_r) &= \operatorname{Var}(\mathbf{x}) - \frac{1}{4}\operatorname{Var}(\mathbf{x}_p) \\
&= 2\left[\mathbf{P}\circ\boldsymbol{\Sigma} + (\mathbf{1}-\mathbf{P})\circ\mathbf{B} + \mathbf{A}\right] - \frac{1}{4}4\left[(\boldsymbol{\Sigma}+\mathbf{B}) + 2\mathbf{A}\right] \\
&= (2\mathbf{P}-1)\circ(\boldsymbol{\Sigma}-\mathbf{B}).
\end{aligned}$$

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 $\mathbf{w}$ and that the GLS weights are the variance-covariance matrix of the random variation in the genotype vector have not been made before.

For the parental PGI coefficient, we obtain

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

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 that 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

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

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,

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

where each coefficient has a GLS interpretation, where the dispersion matrix is the variance-covariance matrix of the random variation in the genotypes. (This matrix is proportional to the LD matrix—the variance-covariance matrix of the total variation in the genotypes—when there is no assortative mating.)

Appendix D. *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 the 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:

$$\text{(D1)} \qquad \frac{N_{\mathrm{X}\cup\mathrm{PGI}}}{N_{\mathrm{X}}} \lesseqgtr \frac{1-\left(R^2_{\mathrm{X}}+R^2_{\mathrm{PGI|X}}\right)}{1-R^2_{\mathrm{X}}}.$$

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 (D1) 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%.