A Wolf in Sheep's Clothing:
Idealisations and the Aims of Polygenic Scores

Davide Serpico

Department of Economics and Management, University of Trento

Interdisciplinary Centre for Ethics & Institute of Philosophy, Jagiellonian University

Studies in History and Philosophy of Science (2023), 102, pp. 72-83

https://doi.org/10.1016/j.shpsa.2023.10.006

Abstract: Research in pharmacogenomics and precision medicine has recently introduced the concept of Polygenic Scores (PGSs), namely, indexes that aggregate the effects that many genetic variants are predicted to have on individual disease risk. The popularity of PGSs is increasing rapidly, but surprisingly little attention has been paid to the idealisations they make about phenotypic development. Indeed, PGSs rely on quantitative genetics models and methods, which involve considerable theoretical assumptions that have been questioned on various grounds. This comes with epistemological and ethical concerns about the use of PGSs in clinical decision-making. In this paper, I investigate to what extent idealisations in genetics models can impact the data gathering and clinical interpretation of genomics findings, particularly the calculation and predictive accuracy of PGSs. Although idealisations are considered ineliminable components of scientific models, they may be legitimate or not depending on the epistemic aims of a model. I thus analyse how various idealisations have been introduced in classical models and progressively readapted throughout the history of genetic theorising. Notably, this process involved important changes in the epistemic purpose of such idealisations, which raises the question of whether they are legitimate in the context of contemporary genomics.

Keywords: Polygenic Scores; GWAS; Quantitative genetics; Genetic prediction; Idealisations; Predictive accuracy.

Introduction

A major focus of precision medicine and pharmacogenomics is the prediction of how individuals' genetic characteristics affect their response to drugs and environmental intervention. Investigation of the genetics of complex diseases figures among the core aims of precision medicine also to prevent the onset of symptoms at early stages of development. Recent work in this area is based on quantitative genetics – the branch of genetics focusing on complex traits – which has recently introduced the concept of Polygenic Score (PGS) and Polygenic Risk Score (PRS).

Since genetic information on complex traits is often difficult to interpret causally, these scores do not aim to provide a better mechanistic understanding of the aetiology of diseases, but rather aim to predict phenotypic outcomes given an individual's genetic profile. More specifically, Polygenic Risk Scores (PRS) aggregate the contribution of many genetic variants to the *risk* of developing binary (yes/no) diseases; PRS are thus more often cited in the medical literature investigating disease risk. Polygenic Scores (PGS), instead, represent the role of such variants in between-individual variability in quantitative traits like human stature, body-mass index (BMI), educational attainment (EA), and the Intelligent Quotient (IQ). Since the development of these traits is not well captured by the notion of risk, PGS are more often cited in the context of this type of traits.[1]

The popularity of PGSs is increasing rapidly, especially in the behavioural sciences.[2] Although the methods for calculating them are relatively new and guidelines for their use are still under development (Kullo et al. 2022), some applications are already in practice, and more are about to come. Private healthcare and direct-to-consumer companies are marketing them as part of in vitro fertilisation in countries such as the US and Japan (Nature Editorial 2022; NHGRI Online Resource). Individuals' genetic data can also be uploaded to online tools to calculate the score for specific diseases, such as type 2 diabetes, breast cancer, and cardiovascular diseases. The future potential of PGSs is uncertain, but scholars have argued that they will eventually become an invaluable tool in medical and social sciences (for an enthusiastic defence, see Plomin

[1] To a certain extent, the distinction is however historical and contingent: one can consider either the PRS or the PGS of a trait depending on whether one conceptualises such trait as binary (qualitative) or continuous (quantitative). For example, schizophrenia has been conceptualised in both senses in different areas and periods of psychiatry genetics (see Serpico 2020). In this paper, I shall use the term *PGS* to denote both PRS and PGS.

[2] A bibliometric research on the database of Web of Science (WoS) on the query [TS=((“polygenic score*” OR “polygenic risk score*”))] reveals that, among about 5.350 research outputs, more than 4.400 have been published between 2019 and 2023 (up to July 2023). In terms of research areas, more than 2.000 results are in the behavioural and brain sciences (e.g., psychiatry, neuroscience, neurology, psychology). This research can be replicated through WoS Advanced Search (https://www.webofscience.com/wos/woscc/advanced-search).

2019). For instance, PGSs would predict the development of personality and educational attainment to inform personalised education (Wang et al. 2022), but it is unclear whether PGSs will bring substantial benefits in psychiatry (Giangrande et al. 2022).

While many discussions revolve around the predictive power of PGSs, little attention has been paid to the *idealisations* they involve about phenotypic development and to the relationship between idealisations, descriptive accuracy, and prediction. Broadly speaking, idealisations are theoretical assumptions that deliberately misrepresent a system or a phenomenon for some epistemic or practical purpose.[3] In this paper, I focus on limitations of PGSs deriving from idealisations in quantitative genetics models that can generate uncertainty in data gathering and clinical interpretation of genomics findings. What types of idealisations does quantitative genetics introduce as regards pathological development? To what extent, if any, do such idealisations impair the predictive capability of PGSs?

Addressing these questions will promote a better understanding of the role of idealisations quantitative genetics, with extensive implications in terms of an ethically sensitive use of PGSs. This analysis will also help contextualise methodological problems in contemporary genomics within a broader theoretical perspective, according to which much of the uncertainty relating to the use of PGSs depends on their reliance on highly idealised models of the genetic architecture of complex traits.

Indeed, as in many -omics areas of biology, where scientists are dealing with an outstanding degree of complexity, relevant epistemological and ethical questions about contemporary genomics can be easily obscured. Moreover, PGSs derive from a branch of genetics where many genes are thought to be involved in the development of complex traits. This sounds immediately 'more desirable' (both ethically and epistemologically) than the reduction of individual differences to single genes, as with the classical notion of Mendelian trait.[4] However, quantitative genetics involves a core of methodological and theoretical assumptions that have been criticised as questionable at the very least, if not strongly misleading.

Models and data in contemporary quantitative genetics draw on three main sources: classical additive models (Fisher 1918; Falconer 1965), which describe the combined additive effects of thousands of alleles on complex phenotypes; heritability analyses, which provide evidence that

[3] Here, with Levy (2021), I consider idealisations as misrepresentations of some aspect of a phenomenon, in contrast to abstractions, which involve the level of details in a scientific representation or model (see also Portides 2021 and §2 below).

[4] Indeed, quantitative genetics does not imply any straightforward reduction of phenotypes to single genes and seem to prevent popular discourses about the 'gene for *x*,' where *x* is any complex trait (see Kendler 2006; Rutter 2006; Sarkar 1998).

variation in every human trait has a genetic basis; and genome-wide association studies (GWAS), which aim to identify genetic variants statistically associated to phenotypic variation. In the last decades, there have been extensive debates on all such facets of quantitative genetics (Downes & Matthews 2019; Lewontin 1974; Kempthorne 1978; Nelson et al. 2013; Serpico et al. 2023), which come with concerns on the use of PGSs in medical explanation and decision-making (Baverstock 2019; De La Vega & Bustamante 2018). These criticisms lie at the core of a long-standing divide between quantitative genetics and other areas of biology, such as epigenetics and developmental biology: while the former has traditionally focused on statistical genetic effects, models from the latter are more causally oriented and traditionally considered 'less idealised' or more descriptively accurate due to their reliance on mechanistic explanations. Attempts to connect statistical and causal explanations have been made (Lu & Bourrat 2022; Tabery 2014), but the gap between the two is still unbridged (Matthews 2022; Matthews & Turkheimer 2021, 2022; Kaplan & Turkheimer 2021).

Limitations of quantitative genetics methods have potential reflections not only on the estimation of PGSs, but also on the reliability of the clinical decisions that they would recommend, with impactful ethical implications, too. In the context of genetic screening, future parents may make decisions based on evidence from these studies. Genetic tests are also directly available to consumers, who can get information about their genotype without competent supervision, with potentially adverse psychological, ethical, and legal consequences (see Wang et al. 2022: 309). Baverstock (2019), for instance, argued that misrepresentation of PGSs may lead to misplaced confidence in low scores – with the risk that a disease is not diagnosed – or misplaced confidence in high scores – which can lead to unnecessary treatment (biopsies, operations) or unwarranted reproductive decisions. Burt (2022), as another example, invited special caution about the use of PGSs in the social sciences. In the context of prenatal screening, PGSs "could trigger the unnecessary destruction of viable embryos or induce women to undergo extra cycles of ovarian stimulation to collect more oocytes" (Nature Editorial 2022).

Below is a detailed structure of the paper.

In §1, I outline how PGSs are calculated based on GWAS data. Then, I summarise major limitations in standard GWAS that are source of both epistemic and ethical uncertainty; I will suggest, however, that some of such limitations may have little relevance for the PGSs' predictive capability.

In §2, I explain in what sense idealisations in genetic models can impact the calculation, interpretation, and use of PGSs. Idealisations are usually considered ineliminable components of scientific models due to a variety of epistemic benefits. For instance, they may enhance our

understanding of complex phenomena by focusing only on certain aspects of such phenomena, thus having heuristic or explanatory purposes, or help exclude irrelevant correlations, simplifying statistical or mathematical analyses. Nevertheless, scholars agree that an idealisation may be legitimate or not depending on the *epistemic aims* of a model. Since the main aim of PGSs is to *predict* development and individual genetic risk, I mostly frame the discussion in terms of how idealisations could impact the predictive accuracy of PGSs. However, I also argue that prediction is not the only aim of PGSs: within the vision of genomics-based precision medicine, the key aim of PGSs is to devise personalised therapeutic and environmental interventions. This suggests that the legitimacy of PGSs idealisations should consider factors beyond predictive accuracy.

In §3, I focus on three types of idealisations in quantitative genetics models. First, each genetic variant has an additive effect on phenotypic variation (the *additivity idealisation*). Second, genetic risk for most complex traits is normally distributed (the *normality idealisation*). Third, variation at the level of single nucleotides represents the appropriate level of complexity to investigate genetic variability in complex traits (the *SNP-causality idealisation*). Drawing on the existing historical and epistemological literature, I analyse how each of these idealisations has been originally introduced in classical models and inherited by contemporary genomics throughout the history of genetic theorising. Notably, this process involved a progressive re-adaptation of the idealisations above and, crucially, of their epistemic purpose. I thus evaluate such idealisations in terms of whether they are legitimate given the epistemic aims of contemporary genomics, particularly the calculation and practical use of PGSs.

Although much of this analysis intersects with current debates on the role of scientific models in data gathering and interpretation, it is mostly motivated by the need of an accurate assessment of what contemporary quantitative genetics can tell us about phenotypic development, particularly complex medical traits. Indeed, each of the three idealisations above reflects precise theoretical and methodological choices about how to best describe the biology of living beings, with extensive practical consequences. Throughout §3 and the Conclusions, I will explore the clinical implications – in terms of prediction and treatment – of making such choices. This includes, for instance, the *inclusion* versus the *exclusion* of non-additive genetic effects from genetics models; the assumption that most complex traits have a quantitative genetic architecture or not (that is, if they are quantitative traits or are just *operationalised* as such); and the question of whether the analysis of single-nucleotide genetic variants can help us understand inter-individual phenotypic variability.

1. What is a Polygenic Score?

Polygenic Scores express, in one single index, the cumulative effects that multiple genetic variants are predicted to have on individual phenotypes. Statistical associations between genetic and phenotypic variation are established through GWAS, a method that tests up to millions of genetic variants in thousands of individuals in case-control designs without prior hypotheses about how such genetic variants may affect phenotypic variation.[5]

The variants that are investigated by GWAS are called single-nucleotide polymorphisms (SNPs), that is, mutations in single DNA base pairs that are common in most target populations (allele frequency > 1%) and that occur normally throughout a person's DNA (NIH National Library of Medicine 2022). At the phenotype level, the variability under the lenses of GWAS usually involves traits such as human stature, cancer, diabetes, obesity, IQ, and major depression, where many genetic variants are thought to be involved. In short, if compared to more classical forms of genetic prediction aimed at identifying highly penetrant genes of large effect size, PGSs focus on the impact of many genetic variants with small effect on individual differences in complex traits.

The primary output of GWAS are lists of p-values and effect sizes of all the tested genetic variants with variation in the phenotype under investigation. Further analyses are then needed to interpret these data and determine the most likely causal variants (for a methodological review, see Uffelmann et al. 2021). In most cases – especially in psychological traits – each SNP identified through GWAS accounts for only a small portion of phenotypic variance, usually less than 0.1%, and for this reason, geneticists need huge samples to identify significant associations. For instance, the effect size of SNPs associated with educational attainment corresponds to 1.7 weeks of schooling per allele (Lee et al. 2018).

This sort of finding is believed to have confirmed classical additive models developed at the dawn of modern genetics (Fisher 1918; Mather 1941, 1943). According to such models, individual differences in complex traits are due to many independent alleles (hundreds, thousands) that have an additive effect on the phenotype. Quantitative traits differ from Mendelian traits, where discontinuous phenotypic differences are due to single, highly penetrant alleles: individual differences in quantitative traits are thought to be determined by the *average value* that many alleles bring about at the phenotype level in different individuals (Chabris et al. 2015).

[5] In this sense, GWAS differ from classical candidate-gene studies, where the identification of genes depends on considerations on their functional role.

GWAS allowed scholars to successfully identify SNPs involved in traits like obesity, cancer, and autoimmune diseases, which is hoped to eventually reveal the genetic basis of complex traits (Marees et al. 2017). Due to limitations that I will discuss shortly, the data obtained through GWAS are today valued for their *predictive potential*, regardless of knowledge about the causal role of SNPs in development, and it is here that PGSs come into play: the calculation of these scores takes the available information on which SNPs are more present in individuals with a given disease and summarise this information into single scores that aim to predict individual genetic risk for such disease.

The construction process of a PGS requires several steps and methodological choices. For our purposes, it is worth highlighting a few aspects (for further details, see Burt 2022; Fries 2020; Janssens 2019; Marees et al. 2017; Uffelmann et al. 2021; Wang et al. 2022; Woodward & Kendler 2023).

First, a PGS is based on data from a previous GWAS, the *target* or *discovery sample*: the number of risk alleles of an individual needs to be weighted by the effect size of each variant detected by the target GWAS, and the resulting products are summed up across all genetic loci. Second, not every SNPs identified in the discovery sample need to be included in a PGS: there is rather a *selection phase* of which variants will constitute the score, which depends on what p-value thresholds are chosen for the score's calculation. Such thresholds are manipulated, for instance, to exclude SNPs that have little statistical evidence for genome-wide association. The predictive accuracy of PGSs generated by different p-value thresholds is assessed by comparing the PGS's phenotypic prediction with the phenotypic information available, and only the SNPs that have higher predictive accuracy are retained.

1.1 Epistemological and Ethical Questions

The reliance of PGSs on quantitative genetics invite new types of epistemological and ethical challenges.

An often-cited issue is that PGSs are vulnerable to stratification biases (e.g., geographic, ancestral, and/or socioeconomic confounders): there is a relevant risk that the associations identified through GWAS are due to indirect genetic effects mediated by environmental heterogeneity in the samples (Trejo & Domingue 2018; Woodward & Kendler 2023). This connects to ethical questions such as whether PGSs could exacerbate existing health inequities given that most GWAS are conducted on participants of European ancestry (Martin et al., 2019; Nature Editorial 2022). In other words, genomics studies have a "strong Eurocentric bias" (Giangrande et al.

2022), which implies that PGSs may not be transferable from a population to another (Johnston & Matthews 2022; Palk et al 2019).

The ethical significance of PGSs, however, is far more wide-ranging than this. Although genetic information is believed to have potential relevance for prevention and treatment, there is a question of *what* type of information should be pursued through genetic testing and *how* such information could be communicated.

First, as regards the 'what' question, epistemic and ethical uncertainty can arise in case of information overload. According to Bayefsky and Berkman (2021), reproductive decision may need limitation in access to fetal genetic information: indeed, unrestricted access to immense amounts of information, including genetic variants with uncertain phenotypic significance, would detract from autonomous decision-making. For instance,

> it will be increasingly unlikely that any fetus will be entirely free from potentially harmful gene variants. There will be 'no more perfect babies,' and parents who pursue expanded testing will be forced to come to terms with their children's genetic risks or choose to terminate the pregnancy and roll the genetic dice once again (Bayefsky and Berkman 2021: 3).

It is thus important to consider the psychological burden of the parents in making decisions based on uncertain information from GWAS and PGS calculation. This requires guidance regarding testing for not just harmful diseases, but also learning and cognitive disabilities and non-medical traits, which may not turn out to be not immediately useful for decision-making processes.[6] Equally important would be for individuals to develop an understating of the many ethical issues at stake, e.g., a child's right to an open future, the risk of implementing eugenic regulatory practices, and disability rights (for a discussion, see Bayefsky & Berkman 2021).

Second, there is the problem of *how* to communicate information about risk. Lewis et al. (2021), for instance, identified three potential representations of risk assessment through PGSs, each of which with different advantages, disadvantages, and applications. Moreover, scholars have extensively discussed how to understand the role of genetic variants with unknown biological meaning and the impact of 'data tsunami' in medical communication. Indeed, like other -omics technologies, GWAS generate data that are not easily translatable into greater explanatory details. In particular, the small effect size of SNPs comes with notoriously unclear biological

[6] There is, in fact, a sense in which genetic characteristics can be assimilated into an individual's "sense of self" or personal identity (Palk et al. 2019: 8). Relatedly, researchers and individuals diagnosed with a mental condition disagree about whether psychiatric diagnoses should be seen as harmful, humanizing, or liberating (Botha et al. 2020; Kenny et al. 2016). This suggests that genetic information on future pathological development may not have straightforward implications for clinical decisions.

insights (Uffelmann et al. 2021), and it is debated whether SNPs data can have any causal value and potential for treatment (Craver et al. 2020; Kaplan & Turkheimer 2021; Oftedal 2022; Turkheimer 2011; Woodward & Kendler 2023). For instance, Giangrande et al. noted that GWAS helped to prioritise research on alleles involved in various psychiatric disorders but uncovering causal mechanisms "would require experimental studies that either are in nascent stages or are not yet possible" (2022: 14).[7]

The SNPs' small effect size connects to another major issue of GWAS known as the Missing Heritability Problem (MHP). The *missing heritability* is the gap between the heritability of a trait estimated through classical family studies (h^2_{TWIN}) and the heritability accounted for by the SNPs detected through GWAS (h^2_{GWAS}) (Matthews & Turkheimer 2021, 2022).[8] For instance, an impressive study of schizophrenia (37,000 subjects and 113,000 controls) identified 128 SNPs that accounted for about 3.4% of the estimated genetic risk (Schizophrenia Working Group 2014). The MHP arises as the SNPs associated with variation in schizophrenia cannot account for the totality of the heritability of the trait: the h^2_{TWIN} of schizophrenia is about 80%, but the SNPs identified account for just a small percentage.

Uncertainty about the causal role of a genetic variant in development can have unpredictable consequences in the selection of embryos (see Turley et al. 2021). As Wang and colleagues warn,

> a previous study reported that embryo selection based on higher polygenic scores for educational attainment would increase the risk of bipolar disorder by 16% from an absolute risk of 1% to 1.16%. The magnitude of this unintended consequence depends on many parameters, including heritability, genetic correlation, PRS predictive performance, the number of embryos for selection, and the prevalence of the traits in the population (2022: 302).

Overall, PGSs information comes with clinical, epistemological, and ethical uncertainty. Even physicians and practitioners may be uncertain about what information is 'certain' or 'uncertain'

[7] More generally on making causal claims based on quantitative genetics methods, Madole & Harden (2022) have recently argued that the genetic causes of complex traits (particularly behaviours) are to be intended as *shallow* causes: many variants from across the genome relate statistically to phenotypic outcomes, but they are non-unitary (they operate within intricate causal systems), non-uniform (produce heterogeneous effects across individuals), and not immediately explanatory in mechanistic terms. For example, individuals diverge at many points along the causal pathway from genes to behaviour, beginning already at the gene-function level; moreover, the probability that single genes matter for phenotypic variation depends on individual-level environmental exposures.

[8] Matthews and Turkheimer take into account three independent facets of the MHP: the numerical gap, the mechanism gap, and the prediction gap. The authors argue that the MHP, rather than being a merely technical problem, depends on profound theoretical differences between traditional and molecular methods in quantitative genetics.

and how to communicate this. It is thus unsurprising that the current literature is so deeply concerned with the clinical potential of PGSs.

The raising connection between quantitative genetics, PGSs, reproductive, and clinical decision-making leads us to the main target of this article: in what sense idealisations in quantitative genetics may impact the predictive accuracy, calculation, and use of PGSs?

It is worth stressing that epistemological and methodological limitations in GWAS may not affect the predictive power of PGSs. For instance, as I discuss below, the limited mechanistic information of SNPs may be irrelevant for predictive purposes since prediction is based on the aggregation of statistical effects of SNPs, rather than on understanding their causal role. By contrast, the MHP may be relevant for PGSs because the predictive power of a PGS is limited by the amount of variance that the identified SNPs can explain.

2. How Predictive are Polygenic Scores?

Recent theorising on scientific modelling converges on the idea that scientific models typically – and deliberately – involve idealisations that misrepresent or distort some characteristics of the system they depict (Levy 2021; Portides 2021). Not only idealisations are somewhat inevitable components of scientific models but are, in many cases, *desirable* means to achieve certain aims. Idealisations are thus often connected with different aims of scientific research: for instance, they may enhance our understanding of complex phenomena, thus having heuristic or explanatory purposes, or help exclude minor correlations that are considered irrelevant, simplifying statistical or mathematical analyses (Brigandt 2013; Jacquart et al. 2022; Strevens 2021; Potochnik 2020; Weisberg 2012).

This also implies that complete descriptive accuracy may not be required for a model to be explanatorily useful in the context of some reasonably well-constrained research programme. Indeed, the role of idealisations should be evaluated in the context of the specific epistemic and practical goals of a given model. According to Brigandt (2013), an idealisation can be legitimate "if the model ignores only those empirical details that do not contribute to the explanatory aim." More recently, Strevens (2021) provided a framework to assess whether an idealisation is safe to make, that is, whether it impairs a model's ability to achieve its goals: according to this framework, the safety of an idealisation depends on whether its inclusion/exclusion *makes a difference* to the epistemic aim of the model.

Since the main epistemic aim of PGSs is prediction, the question of how idealisations can impact PGSs may need to be mostly framed in terms of *empirical accuracy* (Camacho 2021; Hitchcock & Sober, 2004; Van Fraassen 1980). On this view, the possibility of making successful predictions with PGSs would be more important than their ability to capture 'real' causal relationships between the genotype and the phenotype or explain how the phenotype develops over time – Janssens (2019) puts this in terms of the *criterion validity* of PGSs; Plomin and Von Stumm (2023) argued that "the goal of prediction is to account for as much variance as possible, without regard for explanation"; Matthews & Turkheimer (2022) suggest that bridging the "prediction gap" is somewhat independent from the causal explanation of the genotype-phenotype relationship. Moreover, the connection between causality and prediction of PGSs might be irrelevant in certain areas – e.g., in animal breeding – or given certain objectives (Kaplan & Turkheimer 2021). One may also suppose that some idealisations do not impact polygenic prediction in significant ways. For instance, Briley et al. (2019) suggest that the question of the gene-environment interplay might be unimportant in genetics models as long as such models do not attempt to explain how genetic variation *causes* phenotypic variation.

In short, if just prediction is the aim of PGSs, then the exclusion/inclusion of *some* causal details from genetics models might have negligible effects on a score's predictive accuracy.

How could we know that a highly idealised model is empirically adequate? In the case of models that inform the calculation of PGSs, a plausible test would be to check how much phenotypic variance is accounted for by the genetic variants detected through GWAS that are aggregated in PGSs. If the answer is 'very much' or 'about 100%,' then it might be less interesting to make a philosophical analysis of how idealisations impact the calculation of PGSs – based on the pragmatic assumption that we may not need to know much mechanistic details if PGSs have already proved to be highly predictive. However, to the best of our knowledge, the real answer is that the predictive capability of PGSs is *a matter of degree*, and there is wide disagreement about whether it is adequate for their application in the medical and social sciences.

Palk et al. (2019: 4), for example, explain that PGSs are currently able to explain between 1 and 15% of the variance between cases and controls, and argue that such percentages should not be underestimated. Wang et al. (2022) consider in more detail the areas of medicine where PGSs have demonstrated better performance – breast and prostate cancer, type 1 and 2 diabetes, and cardiovascular diseases. Plomin and Von Stumm (2023) reviewed recent increases in the PGS's predictive accuracy for behavioural traits:

> [PGSs] for schizophrenia, which predicted up to 3% of the liability variance in 2009, can now predict 6%. [PGSs] can predict 2% of the liability variance for major depressive disorder, 5% for bipolar

> disorder, 3% for neuroticism, 6% for attention deficit hyperactivity disorder and 10% for externalising behaviours. […] variance predicted by [PGSs] is 7% for general cognitive ability (intelligence), 11% for years of schooling (educational attainment (EA)) and 15% for tested school performance at age 16, which is the most predictive [PGS] in the behavioural sciences.

Various scholars have nonetheless expressed concerns about the performance of PGSs. According to Janssens (2019: 144), the predictive validity of PGSs is modest except when they include one or more SNPs that have a strong impact on disease risk. Janssens & Joyner (2019) argue that the aggregation of millions of SNPs (with close to zero statistical effects) into a single score may be unsuited to informing about the risk of developing a disease.[9] Mostafavi et al. (2020) and Wang et al. (2022) point out various limitations in the performance of PGSs, including inconsistency across studies due to ancestry, population stratification, and trait-specific genetic architectures. Turkheimer (2022) discusses differences in the PGSs' predictive capability in different study designs of educational attainment. Burt (2022: 24), too, disregards the predictive utility of PGSs in education.

I do not aim to take a firm position on whether PGSs are predictive enough: the above disagreements suggest already that further investigation of the epistemological basis of polygenic prediction would be beneficial to secure that idealisations carried out by PGSs are legitimate given their predictive aims. However, there is more than just prediction in PGSs.

2.1. Polygenic Scores Have Many Aims

One may defend an instrumentalist perspective according to which, among the various goals that can motivate a scientific model, predictive accuracy is all that matters for PGSs. However, the relationship between idealisations and descriptive accuracy may be somewhat relevant in the case of PGSs. Below, I make the case that a solely predictive view of PGSs – detached from explanatory aspects – would impair the very calculation, interpretation, and application of PGSs.

For a start, we know that the development of a model involves choices as regards *what* to represent and *how* (Harvard et al. 2021; Harvard & Winsberg 2022), and it is here that idealisations usually come along. As I mentioned, such choices depend on what is expected to better fit

[9] Notably, the inclusion of millions of SNPs that do not meaningfully change risk prediction may have consequences in medical communication: the score would suggest that such SNPs do in fact matter in the development of a disease and that scientists know how SNPs affect diseases. However, neither of these conclusions would be correct: only 10s to 100s SNPs have been reliably associated with any disease, and most of them have no clear mechanistic or causal relationship with phenotypic outcomes. Khera et al. (2018) claimed that PGSs using millions of SNPs outperformed those based on statistically significant SNPs and that the scores identified individuals with risk equivalent to monogenic mutations. Janssens & Joyner (2019) argued that none of these conclusions is warranted.

with the epistemic aims of a model. Even in cases where descriptive accuracy is not the main target, a highly idealised model may be committed to capturing some ontological aspects, such as what entities 'exist' and how they interact to generate a phenomenon. For instance, mechanistic models idealise several details and usually choose specific ways to decompose a system (Craver & Kaplan 2020; Love & Nathan 2015; Piccinini & Craver 2011). This provides a somewhat 'static' and simplified picture of how the components of a system (molecules, cells, etc.) interact with each other but are nonetheless intended to represent real causal processes and interactions.

Something similar may be true for predictive models. As Strevens (2021: 93) explains, predictive models do often 'black-box' some causal elements but, in other cases, they are "causal in character, that is, […] attempt to predict phenomena using a representation of certain causal structures that characteristically produce those phenomena." If so, one may wonder how descriptively inaccurate a model can be without losing some relevant epistemic properties, predictive accuracy included. In the case of PGSs, their prediction certainly has some causal characteristics, that is, they aim to predict phenotypic development from knowledge about the genotype. One can thus agree that their predictive accuracy could in part depend on their ability to capture actual causal interactions between the genotype and the phenotype, such as the role of genotype-environment interactions (G×E) that can mediate genetic effects (I return to this point in §3).[10]

Moreover, the prediction of phenotypic development is far from being the only aim of PGSs: rather, they probably have some *explanatory* and *practical* aims, too.

Starting with explanatory aims: although GWAS do not usually explain how the genotype brings about phenotypic traits, they nonetheless provide data to stratify risk. This is one key aim of PGSs: assessing whether an individual is at the low or high end of a normal distribution of genetic risk (NHGRI Online Resource). For this reason, it might be important that the SNPs aggregated in a PGS are *causally relevant* for the genetic risk of an individual: they should be actual (or plausible) difference makers.

There are also more practical reasons to ask for some degree of descriptive accuracy in genetics models: in genomics-based precision medicine, the calculation of individual scores would ultimately guide clinical decision-making in prevention and personalised treatment. Importantly, even if prediction with PGSs was nearly perfect, this alone could not guide clinical decisions: indeed, to intervene at any level (pharmacological, psychotherapeutic, educational), we need to achieve some reliable causal knowledge. For instance, to prevent the onset of a complex

[10] Woodward & Kendler (2023) have recently made a similar point: "a good PRS is not just predictively successful but must also satisfy the stronger condition that it is predictively successful *because* it tracks underlying genetic influences that are causal."

disease like obesity, we need to know *what* G×E interactions matter (not just that *some* G×E probably matter) to prevent potentially obesogenic environmental factors (for similar considerations, see Briley et al. 2019; Génin & Clerget-Darpoux 2015; this move away from predictive capability was also advised by Janssens & Joyner (2019) concerning genetic consultancy).

To summarise, even if the main epistemic focus of PGSs is prediction, not every type of idealisation in genetic models is, in principle, equally legitimate: while some of them may have little impact on prediction, others may alter a model's capacity of representing causal relationships between genetic and phenotypic variance and, in turn, the PGSs predictive capability. If so, it would be a mistake not to investigate what idealisations may possibly make a difference in polygenic prediction. That said, prediction is probably *not* the only epistemic aim of PGSs: to stratify risk and inform clinical decisions, genetics models need to meet some degree of descriptive accuracy – regardless of the brute predictive power of PGS.

In the next section, I discuss three types of idealisations involved in quantitative genetics models and in the calculation of PGSs. For each idealisation, I analyse questions such as: how does this idealisation enter models and methods? how does it impact the interpretation of the data? does it impair the predictive and descriptive accuracy of PGSs?

3. Idealisations in Polygenic Scores

As I mentioned in the Introduction, the idea that complex traits are highly polygenic is widely uncontroversial and is probably a key reason behind the success of quantitative genetics. There are, however, several corollaries of polygenic models that have instead attracted criticism. Here, I focus on aspects that relate to the following three idealisations: a) each SNP has an *additive effect* on phenotypic variation; b) genetic risk for most complex traits is *normally distributed*; c) variation in single nucleotides (SNPs) represents the *appropriate level of causal complexity* to investigate genetic variability in complex traits.

To highlight the risks of introducing these idealisations in PGSs, I draw on Strevens' (2021) framework to assess the *safety* of idealisations in scientific models, particularly the *logical difference-maker* criterion: an aspect x is a non-difference-maker for the prediction of a target phenomenon y if x can be removed from the model without affecting the model's predictions. By contrast, if x is a difference-maker for prediction, idealising x is not safe. In our case, idealisations in quantitative genetics models are legitimate if they do not affect the calculation of PGSs in such a way as to impair their predictive accuracy and potential use in clinical contexts.

3.1. Classical Models

Idealisations in contemporary GWAS and PGSs derive historically from idealisations made in early quantitative models (Fisher 1918; Falconer 1965; Mather 1941, 1943), which developed both the polygenic theory and the methods to analyse genetic sources of variance that geneticists used for many decades. GWA data and their typical interpretation are tied to the acceptance of such classical models and, at the same time, provide empirical support for such models. It is thus from such classical models that we need to reconstruct the origins of the idealisations involved in the calculation of PGSs.

In the context of early genetics, where most data on genetic causality were gathered in experiments on plant and animal models, the architects of modern genetics were committed to reconciling Mendelian data with the statistical analyses of phenotypes developed by biometricians for the study of complex traits in humans. This required explaining how single genes could generate the continuous and normal distribution of such traits (Downes & Turkheimer 2021). Fisher (1918: 402) argued that the continuous distribution of phenotype values could depend on the involvement of many small and additive genetic effects:

> Let us suppose that the difference caused by a single Mendelian factor is represented in its three phases by the difference of the quantities *a*, *d*, *-a*, and that these phases exist in any population with relative frequency P, 2Q, R where P + 2Q + R = 1. [...] a^2 then is the variance due to this factor, for it is easily seem that when two such factors are combined at random, the mean square deviation from the new mean is equal to the sum of the values of a^2 for the two factors separately. [...] To justify our statement that a^2 is the contribution which a single factor makes to the total variance, it is only necessary to show that when the number of such factors is large the distributions will take the normal form.

It is on this hypothesis that quantitative genetics is based: it is assumed that variation in complex traits is caused by a large (infinite, in Fisher's model) number of alleles (SNPs, in contemporary terms), each of which with (infinitesimally) small effects on the phenotype. Moreover, genetic variation is normally distributed, and, at the population level, the average effects of alleles are reflected linearly on average phenotypic differences. In many mathematical descriptions (e.g., Panse 1940: 104; Purcell 2013: 374), a fixed coefficient is attributed to each allele. For instance, Mather (1943: 39-40) explains:

> With only three polygenes of equal effect, the genotypes AABBcc, AAbbCC and aaBBCC will, for example, give the same phenotype. This phenotype would also characterize the genotypes AaBBcc, AABbcc, AaBbcc, etc., if dominance were the rule, or AABbCc, AaBBCc, and AaBbCC in the absence

> of dominance. […] The allelomorphs designated by small letters are assumed to add nothing to the expression of the character, while each allelomorph designated by a capital letter adds 1 unit. […] As the number of genes involved increases, more phenotypes are possible, and the distribution becomes more nearly continuous […] as observed, for example, in human stature.

As I show below, with the development of genetic epidemiology and molecular biology, geneticists capitalised on these additive models to explain far more than just how the normal distribution of phenotype values could originate from the combination of many independent genes: the model was recruited to analyse heritability (the portion of phenotypic variance accounted for by additive genetic variance) and, later, provided the basis for GWAS aimed at identifying the genes that account for a trait's heritability. Over time, the idealisations made by early geneticists have thus been re-adapted across different types of methods, data, and research contexts. Remarkably, such idealisations were originally introduced in a different explanatory and methodological context to serve specific epistemic aims, which raises the question of whether their re-adaptation is legitimate given the aims of contemporary genomics.

3.2. Additivity

After Fisher defined non-additive effects as the mere "deviation" from the expectation of typical allelic effects (Fisher 1918: 403-408; Phillips 2008), the additivity idealisation was often made to make genetic and environmental variation *more tractable* and exclude non-additive sources of variance that would generate higher complexity. The first step was operationalising phenotypic variance in terms of the additive effect of two, independent sources of variance (*nature* and *nurture*): the total phenotypic variance (V_P) is modelled as the sum of genetic (V_G) and environmental variance (V_E), meaning that phenotypic values depend on an organism's alleles plus additive environmental factors.

This allowed the application of analysis-of-variance techniques (ANOVA) and the estimation of the popular statistics known as narrow-sense heritability (h^2), namely, the proportion of a trait's variance associated with additive genetic variance. In plant and animal breeding, h^2 provided a useful breeding value coefficient to predict how much a population will change over generations (Downes & Turkheimer 2021; Falconer & Mackay 1996; Schaffner 2016).

Eventually, it became routine to make sense of quantitative traits in biochemical terms, translating statistical additive effects into *biochemical* additive effects. For instance, Pierce (2017, chapter 24):

> If the characteristic is polygenic, many different genotypes are possible, several of which may produce the same phenotype. For instance, consider a plant whose height is determined by three loci (A, B, and C), each of which has two alleles. Assume that one allele at each locus (A+, B+, and C+) encodes a plant hormone that causes the plant to grow 1 cm above its baseline height of 10 cm. The other allele at each locus (A−, B−, and C−) does not encode a plant hormone and thus does not contribute to additional height. If we consider only the two alleles at a single locus, 3 genotypes are possible (A+ A+, A+ A−, and A− A−). If all three loci are taken into account, there are a total of 33 = 27 possible multilocus genotypes (A+ A+ B+ B+ C+ C+, A+ A− B+ B+ C+ C+, etc.).

Similar descriptions of biochemical additivity regard enzymatic effects on pea seeds pigmentation, where continuous variation can be explained in terms of different pigments – encoded by different genes – that additively generate colour (for an interpretation of additivity in pea seeds texture, see Jamieson & Radick, 2013: 583-584).

In short, the development of quantitative genetics reflects a transition from a statistical interpretation of additivity to a causal (biochemical, molecular) one. Notably, this transition comes with changes in the epistemic aims of genetics models: on the one hand, Fisher's model aimed at modelling continuous variation under the assumption that phenotypic variation is caused by Mendelian alleles; on the other hand, recent models attempt to explain how the genotype and the phenotype interact in molecular terms. Over time, the necessity to emphasise the pragmatic origin of the additivity idealisation seems to have gradually faded: for instance, Bouchard (2004) claimed that there is no evidence of non-additive genetic effects for IQ and mental disorders; Plomin et al. (2013: 199) then stated that the absence of non-additive variance is very fortunate for the attempts to identify intelligence genes because this allows studying intelligence with statistical methods within a purely quantitative framework.

In contemporary genomics, the additivity idealisation translates into the practice of focusing exclusively on average SNPs' effects on a trait's distribution and into the assumption that SNPs are nearly inter-independent and can be fit additively (Wang et al. 2022: 296). When it comes to the construction of PGSs, relaxing the p-value threshold below the genome-wide significance (§1) seems to depend in part on an endorsement of Fisher's infinitesimal model (see also below).

Crucially, the additivity idealisation underestimates the role of non-additive gene-gene (G×G) and gene-environment interactions (G×E). Both types of interactions concern background effects where the phenotypic outcome of a genetic variant depends on other independent variables (mutations at other loci or environmental exposures). An example of G×G is epistasis, where an individual with a particular combination of alleles can display a phenotype beyond what is expected from the effects of such alleles if they were independent (Carlborg & Haley 2004; Phillips

2008; Webber 2017). This type of interaction can mask dramatically the G-P map even in single-gene diseases (Campbell et al. 2018). Likewise, G×E are known to be an important source of individual differences (Champagne & Mashoodh 2009; Nagpal et al. 2018; Sauce & Matzel 2018; Schaffner 2016; Tabery 2014; Wahlsten 1994).

Although several studies on model organisms – and a few studies on humans – attest these interactions, GWAS provide little evidence of their relevance and, probably for this reason, G×E are rarely addressed in quantitative genetics literature.[11] But this depends in part on methodological limitations: interactions are difficult to capture in non-experimental studies due to their contextual nature, their small effect size and, in the case of G×E, the lack of standardised measures of environmental influences (Assary et al. 2018; Matthews & Turkheimer 2022).[12]

For our analysis, the key aspect is that quantitative genetics models are designed to exclude non-additive sources of variance for the pragmatic reasons outlined above. In this sense, the observation that interactions have a negligible role might be *model-dependent*, meaning that it could depend on how data are gathered and interpreted based on the model.

As Nelson et al. (2013: 673) noted, "the current quantitative genetics framework has a built-in bias against inferring epistasis, making it insufficient for identifying and interpreting epistatic effects." Burt (2022: 19) acknowledged that "evidence for a substantial role of interactionism is lacking; however, the current evidence is primarily based on low resolution tag SNP methodologies [that] have not yet substantiated the importance of gene-gene interactions [but do] not suggest they are not biologically important." If compared with the multitude of quantitative geneticists advocating the negligible role of non-additive phenomena, these are sparse voices. They nonetheless suggest that the calculation of PGSs may be model-dependent to such an extent that it overlooks phenomena that are not predicted by the additive model.

If additivity misrepresents the complexity of the genotype-phenotype-environment relationship and masks the complexity of disease aetiology, this may have major consequences for the predictive accuracy of PGSs and their potential to inform treatment.

[11] For example, Plomin and Von Stumm (2023) only discuss the role of G×E correlations, which represent a statistical association between genetic and environmental variances, rather than causal interactions.

[12] Association studies would need much higher statistical power than regular GWAS to detect such interactions (Dai et al. 2018; Carlborg & Haley 2004; Giangrande et al 2022; Nagpal et al. 2018); however, increasing sample size introduces more heterogeneity and may thus impair the results further (Hellwege et al. 2018; Marchini et al. 2004; Webber 2017). This also connects to the problem of the portability of PGSs across populations: like classical heritability analyses (Block 1995), the effects of SNPs on the phenotype are mediated by context-relative interactions between genetic and environmental variation (Burt 2022; Janssens 2019; Johnston & Matthews 2022; Matthews 2021; Wang et al. 2022). Note that the PGSs predictive accuracy can also differ substantially across groups of similar ancestry (Mostafavi et al 2020).

As regards prediction, Woodward and Kendler (2023) recently noted that it may seem surprising that PGSs are predictively successful given that they combine SNPs information according to a simple additive formula. As I explained in §2, however, the predictive power of PGSs comes in degree and is subject of controversy, so it worth considering whether prediction could be improved by not making the additivity assumption. Our knowledge of G×E does suggest that the additivity assumption might be a difference-maker for prediction. Studies on obesity, for instance, show that PGSs explain different amounts of variance in groups of individuals exposed to different environmental factors such as smoking, physical activity, or sugar consumption (Nagpal et al. 2018: 2; on diabetes, see Génin & Clerget-Darpoux 2015). Furthermore, neglecting interactions can decrease our ability to predict how the presence/absence of a genetic variant will impact other (apparently unrelated) traits. For instance, if one selects an embryo with a lower risk of developing a disease, the embryo may have a higher susceptibility to other conditions that are pleiotropically affected by the excluded genetic variant (Nature 2022: 549). So, should be proved that non-additive phenomena are difference-makers for prediction, then the additivity idealisation would turn out to be *illegitimate*.

As regards intervention, the idealisation arguably impairs our ability to target effectively pathological conditions through environmental intervention: by default, PGSs data will not include information regarding how pathologies could interact with the environment – rather, they tend to obscure the importance of non-genetic aspects. If we aim to use PGSs in real-world clinical scenarios, we need to understand the context-dependent role of genes as well as potential sources of stratification biases (literature on G×E and heterogeneous response to treatment testifies this clearly, see Farzan et al. 2018; Kersten & Koppelman 2017; Chang et al. 2015; Keers & Uher 2012; Pedersen 2017; Uffelmann et al. 2021). Plomin and Von Stumm (2023: 51) argued that "the issue of whether population stratification confounds polygenic score prediction in a particular population is separate from the ability of polygenic scores to predict in different populations". However, as I argued in §2.1, intervention cannot be completely detached from mechanistic explanation: since polygenic prediction would ultimately guide personalised intervention (e.g., pharmacological, psychotherapeutic, educational), we need to know what G×E interactions matter for individual risk, that is, why – and in what contexts – a genetic variant can lead an individual to develop a disease or not.

I am not suggesting that additive genetic effects do not exist at all, but I am making the case that a non-statistical interpretation of additivity makes probably sense in the case of simple biochemical interactions, where the phenotype is the result of a biochemical cascade of pigments. And these are, in fact, the cases where the additivity idealisation was originally introduced (see

above). However, additivity is often taken *literally*, as a 'concrete' aspect of the G-P relationship, even though this does not match our knowledge of biological systems characterised by multiplicative interactions between molecular, cellular, and environmental variables.

3.3. Normality

A second idealisation that derives from classical models is that *genetic risk is normally distributed*. As in the case of the additivity idealisation, normality depends on theoretical choices that have been introduced for pragmatic reasons, such as the aim to estimate quantitative genetic parameters in animal breeding. In fact, Fisher already posed some constraints on the model's applicability: he aimed to characterise variation *among relatives* and to show that, "under the infinitesimal model, the distribution of genetic components within families remains normal, with variance that evolves in a way that is entirely determined by relatedness" (Barton et al. 2017: 50). As Fisher (1918: 400) puts it:

> Speaking always of normal populations, when the coefficient of correlation between father and son, in stature let us say, is r, it follows that for the group of sons of fathers of any given height the variance is a fraction, $1 - r^2$, of the variance of sons in general.

After Fisher, the normality idealisation had also much potential for the application of variance-partitioning studies to yes/no traits, like pathologies that are either diagnosed or not. In 1965, Falconer extended Fisher's infinitesimal model assuming that genetic liability is polygenic and continuously distributed even for these traits, thus reducing *discontinuity* at the phenotype level to *continuity* at the genotypic level. More recently, geneticists like Plomin et al. (2009) and Knopik et al. (2017) argued that continuity is not only observed at the genotype level, but also at the phenotype level (e.g., each of us feels occasionally more or less 'sad', and for some of us this translates into clinical conditions like major depression).

In PGSs, normality comes as a heuristic to stratify genetic risk: starting from the mathematical models above – where a fixed coefficient is attributed to each allele – PGSs aggregate the additive effects of GWAS-weighted allele risk and normalise such risk to obtain scores with a mean of 0 and standard deviation of 1 (Lewis & Vassos 2017). As the Broad Institute Online Resource explains:

> Some people will have a higher or lower score, depending on their number of risk-increasing or risk-decreasing variants, and the magnitude of impact of each variant. People with an approximately even number of risk-increasing and risk-decreasing variants are at an average risk of disease based on their

> genetics. People with more risk-increasing variants are at an increased risk of disease based on their genetics. People with more risk-decreasing variants are at a decreased risk of disease based on their genetics.

Essentially, this provides a "human equivalent of the 'breeding value' in selective plant and animal breeding" or a single quantitative measure of genetic risk (Burt 2022: 10) that is convenient for many reasons. Among them, normalised PGSs allow one to determine an individual's position on the distribution of risk and whether she crosses a sufficiently high threshold (e.g., beyond the 5% of the population-wide risk) and thus requires clinical attention (Palk et al. 2019).

As I mentioned above, Fisher's model does not imply that a trait's distribution *in the whole population* (which is the target of GWAS) is *normal.* But there are three specific reasons to think that genetic risk is not distributed normally as classical models predicted.

First, while single-nucleotide variants might be distributed normally, this is probably not the case for other types of genetic variants, particularly rare and copy-number variants (the role of which is becoming increasingly clear in mental conditions, see Giangrande et al. 2022). PGSs exclude systematically the role of non-common variants that often account for a larger proportion of phenotypic variance: indeed, for the most part, these variants are difficult to capture through GWAS as they may be found only in small subgroups – so that they are not sufficiently represented in GWA large cohorts – and have imperfect penetrance (Baverstock 2019; Burt 2022; Fries 2020; Génin 2020; Giangrande et al. 2022; McClellan & King 2010; Zaidi & Mathieson 2020). Thus, the normality idealisation risks masking the heterogeneity of genetic effects on population variance.

Second, it should be noted that, beyond theoretical models, the main piece of evidence for the normal distribution of genetic risk is the presence of disease-associated SNPs in non-clinical populations (e.g., Owen et al. 2007; Riglin et al. 2016; Robinson et al. 2016). These data are interpreted as evidence that the genetic basis of complex traits, including behaviours, is continuously distributed (e.g., for Plomin et al., 2009: 877, evidence of polygenicity seems enough to conclude that a trait is quantitative). But in fact, GWAS only suggest that complex traits are highly polygenic and that their distribution does not follow simple patterns (i.e., there is no specific genetic variant that only 'diseased individuals' have and 'healthy' individuals don't). In other words, these data do not imply that there are individuals with an average number of 'average-value' alleles (e.g., alleles that bring 0 as a phenotypic value in the mathematical models above) and that fewer individuals have a smaller number of 'good-value alleles' (+1) and 'bad-value alleles' (-1) (note that recent theories account for complex traits beyond a quantitative characterisation, see Boyle et al. 2017; Serpico 2020; Serpico & Petrolini 2023). This is a case where the

normality idealisation forces the interpretation of the data under a Fisherian framework even if the data have other plausible interpretations.

A final aspect is that the normality *of phenotypes* was, in classical models, a key reason to assume normality *at the genetic level*. However, this depends on whether a phenotypic trait is *actually* quantitative or is rather *operationalised* as such. As I mentioned above, classical models have been intended to provide a genetic account of physical traits like human stature (expressed in cm), body-fat content (expressed in BMI values), and behavioural traits like intelligence (operationalised through IQ scores). The normality idealisation was suggested by the observation that phenotype values vary continuously in populations, rather than by empirical data on the *actual* distribution of genetic factors. Fisher himself had no empirical reason to believe that genetic variation is continuous: he aimed to provide a plausible – though idealised – explanation of how underlying genetic variability can generate phenotypic continuity. On this view, values in quantitative traits are assumed to be normally distributed *because genetic variation is assumed to be distributed as such*.

This can be further clarified by Falconer's model, which introduced a theoretical concept called *liability* denoting "an underlying gradation of some attribute immediately related to the causation of the disease" (1965: 52). Notably, Falconer did not make any "unwarranted assumption about the real nature of the liability." Such a notion "simply specifies that in order to express the degree of liability we shall choose a scale of measurement which, if we could measure the liability, would yield a normal distribution" (1965: 53). In this sense, the normality idealisation was not made to describe the 'real' relationship between genetic risk and phenotypic variation, but rather to provide a quantitative interpretation of the genetics of discontinuous traits that would allow for the application of ANOVA and heritability analyses to yes/no traits like diseases.

The normal distribution of genetic liability is, in other words, predicted or postulated by such models, rather than simply observed. While a quantitative description makes sense in the case of human stature, quantitative variation in other traits can depend on how we operationalise and normalise them. For instance, operationalising body-fat content through BMI (as well as intelligence through IQ scores) may be convenient to analyse individual differences: the data get more tractable and can be inspected through ANOVA. However, this operationalisation involves precise methodological choices about how to normalise individual raw values/scores and ultimately requires reducing individual variation in such complex traits to unidimensional metrics. One should not confuse the phenotypic continuity generated by our measurement methods with continuity in a stronger, ontological sense (Koi 2021; Meehl 1992, 1999; Michell 1997, 2012; Hibberd 2014; Serpico & Borghini 2021; Serpico 2018, 2020; Serpico & Petrolini 2023).

If we begin to cast doubts on the very possibility that genetic risk is normally distributed as quantitative genetics have initially postulated – and repeatedly assumed – the attempt to predict the 'position' of an individual on the 'normal distribution of risk' through PGSs loses part of its appeal. Moreover, an immediate implication of the normality idealisation is that PGSs would mask heterogeneity effects at different positions of the distribution of risk. Indeed, together with the additivity idealisation (§3.2), it is assumed that, on average, additive genetic effects will be similar for individuals at the low and high end of the distribution. This prevents accounting for individual differences in response to environmental influences (for similar considerations, see Wehby 2018).

The normality idealisation is probably *not* a difference-maker for prediction, in the sense that its inclusion in genetic models may not affect the predictive power of a PGS. However, this idealisation justifies the very idea of a quantitative score of genetic liability: only if we embrace the view that genetic risk is normally distributed does polygenic prediction make sense. If we instead exclude this idealisation from genetics models, different types of genetic prediction – not involving the normalised effects of GWAS allele risks – happen to make much more sense.

3.4. SNPs Causality

There is extensive literature about whether GWAS data can be interpreted causally. But less explored is the question of whether single-nucleotide variants (SNVs), the simplest type of genetic variation, represent the appropriate level of complexity to analyse genetic causes. The fundamental assumption of contemporary genomics is that the effects of these genetic variants can be observed at the phenotype level because *these* are the variants that, ultimately, generate phenotypic variation. Or they are just easier to detect. Either way, these variants are investigated in case-control studies where the DNA of a population with a trait is statistically compared with the DNA of a population without the trait: the expectation is that the SNPs that are causally relevant for that pathological phenotype will be more common in the case group. What if, however, GWAS have significantly idealised the role of SNVs in development?

Let us return once again to classical models. Fisher speaks of *Mendelian factors* which, back then, did not have a clear material nature. With the development of molecular genetics, genetic variation was characterised in terms of DNA point mutations. It seems reasonable to think that such small mutations would correspond to Fisher's alleles – the infinitesimally small genetic causes that would explain the normal distribution of traits. So, one may think that the transition from classical models to contemporary genetics is a case of successful reduction of one scientific

theory to another: it just turned out that the abstract concept of *allele* used by early geneticists was to be reduced to *SNPs*. On this interpretation, it is also natural to interpret the small effect size of SNPs as a confirmation of classical models, because genetic effects are so small that standard GWAS cannot identify them easily (§1).

Many critics argued that most SNPs may have little relevance for phenotypic variation as they are probably functionally irrelevant (Baverstock 2019; Oftedal 2022; Richardson & Jones 2019). Here, I want to emphasise a different aspect: although the SNPs identified by GWAS may have some functional – yet unknown – role in phenotypic variation, they may not be the genetic causes that we are looking for when we talk about *alleles*. There are in fact various reasons to think that the concept of allele does not correspond to the concept of SNP but can rather reflect genetic variation of different sorts.

To illustrate, let us consider a few cases of single-gene conditions. Sickle-cell anaemia is due to single-nucleotide substitution (A with T) in the gene encoding haemoglobin. In the resulting peptide of individuals carrying the allele, a GAG codon is substituted with a GTG codon, which encodes for a *valine* amino acid at position 6 instead of *glutamate*. This is a case where analysing variation at the SNV level makes perfect sense because a single-nucleotide difference generates the observed phenotypic difference.

However, conditions of this sort constitute a relatively small set of single-gene diseases. More often, such conditions involve mutations at many sites of one gene (not just one mutation) and such mutations are usually not the same in every individual diagnosed with the condition. For instance, hundreds of different mutations are responsible for phenylketonuria (Kronfeldner 2009; Scriver 2007). As another example, most cases of Huntington's disease are caused by multiple copies of the CAG codon (Plomin et al. 2013: 130). That is, what is similar in different individuals is not a single-nucleotide mutation, but rather the impaired functionality of an encoded protein, which can be also determined by factors beyond the DNA sequence (e.g., ranging from cellular to environmental interactions). So, even in relatively 'simple traits' like Mendelian diseases, phenotypic variability can be realised by different mutations and be mediated by background effects.

The key point is that the theoretical concept of *allele* of classical models does not strictly correspond to the concept of *single-nucleotide variant*. Alleles, intended as genetic variants that bring about a protein with impaired functionality, are multiply realised: they can have similar phenotypic effects (e.g., encoding of a dysfunctional protein, disrupting cellular functionality) despite different underlying structures or chemical bases, e.g., they can involve different *types* of variation, such as repeated codons or point mutations (a similar point was raised by Craver et al. 2020: 1090; see also Bourrat 2020; Oftedal 2022).

One consequence is that most SNPs may have little or no phenotypic effect in many individuals. If so, the GWAS strategy of dividing a population into case and control groups can be risky: if the SNP-causality idealisation is inaccurate – and allelic variation is not reducible to single-nucleotide variation – we cannot expect stable correlations between the presence of a disease and the presence of simple variants like SNPs.

While in most GWAS this risk is mitigated given standard requirements regarding the genome-wide threshold of statistical significance for the inclusion of SNPs, PGSs are usually calculated by lowering p-value thresholds to include as many SNPs as possible (§1). Through this methodological choice, the selected p-value threshold of SNPs often is <1, which is below the typical threshold of genome-wide significance. Unsurprisingly, as Burt (2022) and Janssens (2019) warned, PGSs often end up including every available SNPs regardless of their statistical significance in the target GWAS, that is, genome-wide significant ones plus many others that have (arguably) little or indirect connections with variation in the trait of interest.

Crucially, relaxing p-value thresholds may lead to including in a PGS also variants with low penetrance, i.e., variants that have phenotypic effects only in a small number of individuals. For many of such variants, the penetrance might be so low to make them practically irrelevant for prediction. In this sense, the SNP-causality idealisation likely *is* a difference-maker for the PGSs' predictive power: if SNPs are not to be taken as *alleles*, then there is no clear sense in which a PGS based on SNPs information should be able to predict phenotypic development.

To sum up: based on classical models, SNPs have been taken to represent genetic variability at the relevant causal level: the inference was that the alleles in such models correspond to the smaller units of genetic variability. Given the idealised characteristics of such alleles (small additive effects and normal distribution), SNPs investigated through GWAS are clearly natural candidates for the role of the "infinite number of Mendelian factors" that generate the continuous distribution of traits. This conclusion, however, is legitimate only if we take Fisher's model *literally* and try to translate the idea of an "infinite number of Mendelian factors" into a concept from molecular biology.

From what we know about biological systems, SNVs might not be the relevant genetic factors to investigate the impact of genetic variability on variability in complex traits. This is testified by the fact that genetic effects identified through GWAS are so small that the calculation of PGSs requires manipulating p-value thresholds. But what if we are not looking at the right level of complexity?

Conclusions

The calculation of PGSs relies on highly idealised models of the genetic architecture of complex phenotypes that have been developed in the early days of quantitative genetics. In this paper, I explored whether idealisations in such models can impact polygenic prediction and the use of SNPs information in clinical decision-making. I focused on three types of idealisations: a) each allele has an *additive effect* on phenotypic variation; b) genetic risk for most complex traits is *normally distributed*; c) variation in single nucleotides represents *the appropriate level of complexity* to investigate genetic variability in complex traits. Since the PGSs' capability of predicting individual development is still a controversial topic, I argued that the inclusion of some idealisations in genetic models *makes a difference* to the predictive accuracy of PGSs and their ability to capture the relationship between genetic and phenotypic variation.

These idealisations have played a crucial role in the development of quantitative genetics methods, but it is important to stress that their role have changed throughout the historical development of quantitative genetics.

Idealisations in early models deliberately misrepresented the G-P relationship in a way that decreased their descriptive accuracy, but the pragmatic aims of early geneticists seem to have abundantly justified such representational choices. Over time, classical models have been translated into more *literal* models with stronger descriptive aims: GWAS ultimately aims at capturing causal relations between genetic and phenotypic variation to inform medical decision-making. PGSs have similar objectives: some forms of therapeutic or environmental intervention would follow from prediction. The progressive re-adaptation of genetics idealisations risks making contemporary methods incapable of achieving the epistemic and practical aims of genomics-based precision medicine. Since such aims differ from those of the models where the idealisations were originally introduced, quantitative genetics idealisations seem now less legitimate than they probably were in the early days of genetic theorising.

The presence of idealisations in PGSs conflates with the view that GWAS is a hypothesis-free method: at the very least, GWAS is not a *theory-free* method given that they presuppose a model with specific choices regarding *what* to represent and *how*. I suggested that such choices determine both the data that we can reasonably get from GWAS as well as their interpretation.

To illustrate how idealisations can mislead the generation and interpretation of empirical data, let us consider a hypothetical case. Let us say that 90% of people overestimate their driving skills. One could design a psychometric test including various sub-tests assessing individual differences in such a trait; individuals' sub-test performance could then be summarised into a single

score (say, the Proud Driver Quotient, PDQ), and such scores could be normalised to get a normal distribution. Through this procedure, it becomes possible to analyse the involvement of genetic variation in PDQ differences among individuals through analysis-of-variance techniques: first, twin studies could estimate the percentage of variance in PDQ that is associated with genetic variance (h^2_{TWIN}); second, GWAS could search for the SNPs that account for h^2_{TWIN} (getting a h^2_{GWAS} estimation for PDQ). If there is a gap between h^2_{TWIN} and h^2_{GWAS} for this trait, one can manipulate p-value thresholds to include every SNP below the threshold of genome-wide significance into a PGS.

The question is: would this PGS capture the genetic liability of people's attitude to overestimate their driving skills and predict where, on the normal distribution of PDQ scores, an individual will be?

First, if the additivity and SNP-causality idealisations involve aspects that *are difference makers* for prediction, then the exclusion of non-additive and non-common effects would decrease the predictive power of the PGS. We could thus find out that the inclusion of such effects in the calculation of PGSs explains PDQ variance better than their exclusion.

Second, by dropping the SNP-causality idealisation, it may turn out that focusing on different (less common and more complex) types of genetic variants have a higher potential in terms of individual-level prediction of response to environmental and background conditions.

Third, by dropping the normality idealisation, one could realise that the trait 'overestimating one's driving skills' is *not* normally distributed; perhaps, it is not even a quantitative trait in the standard sense: it is only *operationalised* as such. In that case, assuming that genetic liability is normally distributed would be misleading, and the very attempt to position individuals on a Gaussian curve of genetic risk would make little sense.

The main problem is that data and theories in contemporary quantitative genetics are heavily model-dependent. Sometimes, idealisations enter scientific models in subtle ways, and we may easily lose sight of the relationship between models and empirical findings. For this reason, none of the three idealisations above can be easily ruled out until we stick to standard models. For instance, ever since Lewontin's (1974) seminal paper on the difficulty of making causal claims based on the analysis of variance, scholars have never achieved a consensus on whether additivity is 'the norm' or just 'an exception.' It is beyond my aims to discuss potential ways to improve the ability of GWAS to detect non-additive or non-common sources of variation (on rare variants identification through family-based whole-genome sequencing, see Giangrande et al. 2022; Uffelmann et al. 2021). Until such methods become widely available, the calculation of PGSs based

on GWAS should acknowledge the impact of theoretical idealisation on the descriptive and predictive accuracy of PGSs.

Acknowledgments: This work has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 805498). Throughout my research on this topic, I got invaluable feedback from a variety of colleagues. Among them, I wish to thank the members of the Centre for the History and Philosophy of Science at the University of Leeds, where I presented an embryonic version of these ideas back in 2018; the audience and speakers of the symposium "New Historical and Philosophical Perspectives on Quantitative Genetics" (ISHPSSB, Oslo, 2019); Mark Fedyk for his comments on a previous version of this paper; and two anonymous referees for their insightful suggestions.

References

Assary, E., Vincent, J. P., Keers, R., & Pluess, M. (2018). Gene-environment interaction and psychiatric disorders: Review and future directions. *Seminars in Cell & Developmental Biology*: 77: 133-143.

Barton, N. H., Etheridge, A. M., & Véber, A. (2017). The infinitesimal model: Definition, derivation, and implications. *Theoretical population biology*, *118*, 50-73.

Baverstock, K. (2019). Polygenic scores: Are they a public health hazard?. *Progress in biophysics and molecular biology*, *149*, 4-8.

Bayefsky, M. J., & Berkman, B. E. (2022). Implementing expanded prenatal genetic testing: should parents have access to any and all fetal genetic information?. *The American journal of bioethics*, *22*(2), 4-22.

Block, N. (1995). How heritability misleads about race. *Cognition, 56*(2), 99-128.

Botha, M., Dibb, B., & Frost, D. M. (2020). "Autism is me": An Investigation of how Autistic Indi-viduals Make Sense of Autism and Stigma. *Disability & Society, 37*(3), 427-53.

Bouchard Jr, T. J. (2004). Genetic influence on human psychological traits: A survey. *Current directions in psychological science*, *13*(4), 148-151.

Bourrat, P. (2020). Causation and single nucleotide polymorphism heritability. *Philosophy of Science*, *87*(5), 1073-1083.

Boyle, E. A., Li, Y. I., & Pritchard, J. K. (2017). An expanded view of complex traits: from polygenic to omnigenic. *Cell*, *169*(7), 1177-1186.

Brigandt, I. (2013). Explanation in biology: Reduction, pluralism, and explanatory aims. *Science & Education*, *22*(1), 69-91.

Briley, D. A., Livengood, J., Derringer, J., et al. (2019). Interpreting behavior genetic models: seven developmental processes to understand. *Behavior genetics*, *49*(2), 196-210.

Broad Institute Online Resource. Polygenic Scores. http://polygenicscores.org/explained

Burt, C. H. (2022). Challenging the Utility of Polygenic Scores for Social Science: Environmental Confounding, Downward Causation, and Unknown Biology. *Behavioral and Brain Sciences*, 1-36.

Camacho, M. P. (2021). Beyond descriptive accuracy: The central dogma of molecular biology in scientific practice. *Studies in History and Philosophy of Science Part A*, *86*, 20-26.

Campbell, R. F., McGrath, P. T., & Paaby, A. B. (2018). Analysis of epistasis in natural traits using model organisms. *Trends in Genetics*, *34*(11), 883-898.

Carlborg, Ö., & Haley, C. S. (2004). Epistasis: too often neglected in complex trait studies?. *Nature Reviews Genetics*, *5*(8), 618-625.

Chabris, C. F., Lee, J. J., Cesarini, D., et al. (2015). The fourth law of behavior genetics. *Current directions in psychological science*, *24*(4), 304-312.

Champagne, F. A., & Mashoodh, R. (2009). Genes in context: Gene–environment interplay and the origins of individual differences in behavior. *Current Directions in Psychological Science*, *18*(3), 127-131.

Chang, H. S., Won, E., Lee, H. Y., et al. (2015). Association analysis for corticotropin releasing hormone polymorphisms with the risk of major depressive disorder and the response to antidepressants. *Behavioural Brain Research, 292*, 116-124.

Craver, C. F., & Kaplan, D. M. (2020). Are more details better? On the norms of completeness for mechanistic explanations. *The British Journal for the Philosophy of Science*.

Craver, C. F., Dozmorov, M., Reimers, M., & Kendler, K. S. (2020). Gloomy prospects and roller coasters: finding coherence in genome-wide association studies. *Philosophy of Science*, *87*(5), 1084-1095.

Dai, J. Y., Hsu, L., & Kooperberg, C. (2018). Two-stage testing for genome-wide gene-environment interactions. In *Handbook of Statistical Methods for Case-Control Studies* (pp. 459-474). Chapman and Hall/CRC.

De La Vega, F. M., & Bustamante, C. D. (2018). Polygenic risk scores: a biased prediction?. *Genome medicine*, *10*(1), 1-3.

Downes, S. M., & Matthews, L. (2019). Heritability. In E. N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy* (Spring 2020 Edition). https://plato.stanford.edu/archives/spr2020/entries/heredity

Downes, S. M., & Turkheimer, E. (2021). An early history of the heritability coefficient applied to humans (1918–1960). Biological Theory, 17, 126–137.

Falconer, D. S. (1965). The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Annals of Human Genetics*, *29*(1), 51-76.

Falconer, D. S., & Mackay, T. F. C. (1996). *Introduction to quantitative genetics*. Essex: Longman Group.

Farzan, N., Vijverberg, S. J., Kabesch, M., et al. (2018). The use of pharmacogenomics, epigenomics, and transcriptomics to improve childhood asthma management: where do we stand?. *Pediatric pulmonology*, *53*(6), 836-845.

Fisher, R. A. (1918). The Correlation Between Relatives on the Supposition of Mendelian Inheritance. *Transactions of the Royal Society of Edinburgh, 52,* 399-433.

Fries, G. R. (2020). Polygenic risk scores and their potential clinical use in psychiatry: are we there yet?. *Brazilian Journal of Psychiatry*, *42*, 459-460.

Génin, E. (2020). Missing heritability of complex diseases: case solved?. *Human Genetics*, *139*(1), 103-113.

Génin, E., & Clerget-Darpoux, F. (2015). Revisiting the polygenic additive liability model through the example of diabetes mellitus. *Human Heredity, 80*(4), 171-177.

Giangrande, E. J., Weber, R. S., & Turkheimer, E. (2022). What Do We Know About the Genetic Architecture of Psychopathology?. *Annual Review of Clinical Psychology*, *18*, 19-42.

Harvard, S., & Winsberg, E. (2022). The epistemic risk in representation. *Kennedy Institute of Ethics Journal*, *32*(1), 1-31.

Harvard, S., Winsberg, E., Symons, J., & Adibi, A. (2021). Value judgments in a COVID-19 vaccination model: a case study in the need for public involvement in health-oriented modelling. *Social Science & Medicine*, *286*, 114323.

Hellwege, J. N., Keaton, J. M., Giri, A., et al. (2017). Population stratification in genetic association studies. *Current protocols in human genetics*, *95*(1), 1-22.

Hibberd, F. J. (2014). The metaphysical basis of a process psychology. *Journal of Theoretical and Philosophical Psychology*, *34*(3), 161.

Hitchcock, C., & Sober, E. (2004). Prediction versus accommodation and the risk of overfitting. *British Journal for the Philosophy of Science*, *55*(1).

Jacquart, M., Shech, E., & Zach, M. (2022). Idealization, representation, and explanation in the sciences. *Studies in History and Philosophy of Science*.

Jamieson, A., & Radick, G. (2013). Putting Mendel in his place: how curriculum reform in genetics and counterfactual history of science can work together. In K. Kampourakis (Ed.), *The philosophy of biology*. A Companion for Educators (pp. 577-595). Springer.

Janssens, A. C. J. (2019). Validity of polygenic risk scores: are we measuring what we think we are?. *Human molecular genetics*, *28*(R2), R143-R150.

Janssens, A. C. J. W., & Joyner, M. J. (2019). Polygenic risk scores that predict common diseases using millions of single nucleotide polymorphisms: is more, better?. *Clinical chemistry*, *65*(5), 609-611.

Johnston, J., & Matthews, L. J. (2022). Polygenic embryo testing: understated ethics, unclear utility. *Nature Medicine*, *28*(3), 446-448.

Kaplan, J. M., & Turkheimer, E. (2021). Galton's Quincunx: Probabilistic causation in developmental behavior genetics. *Studies in History and Philosophy of Science*, *88*, 60-69.

Keers, R., & Uher, R. (2012). Gene–environment interaction in major depression and antidepressant treatment response. *Current psychiatry reports*, *14*(2), 129-137.

Kempthorne, O. (1978). A biometrics invited paper: Logical, epistemological and statistical aspects of nature-nurture data interpretation. *Biometrics*, 1-23.

Kendler, K. S. (2006). Reflections on the relationship between psychiatric genetics and psychiatric nosology. *American Journal of Psychiatry*, *163*(7), 1138-1146.

Kenny, L., Hattersley, C., Molins, B., et al. (2016). Which Terms Should be Used to Describe Autism? Perspectives from the UK Autism Community. *Autism, 20*(4), 442-62.

Kersten, E. T., & Koppelman, G. H. (2017). Pharmacogenetics of asthma: toward precision medicine. Current opinion in pulmonary medicine, 23(1), 12-20.

Khera, A. V., Chaffin, M., Aragam, K. G., et al. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature genetics*, *50*(9), 1219-1224.

Knopik, V. S., Neiderhiser, J. M., DeFries, J. C., & Plomin, R. (2017). *Behavioral genetics*. *Seventh Edition*. New York: Macmillan.

Koi, P. (2021). Genetics on the neurodiversity spectrum: Genetic, phenotypic and endophenotypic continua in autism and ADHD. *Studies in History and Philosophy of Science*, *89*, 52-62.

Kronfeldner, M. E. (2009). Genetic determinism and the innate-acquired distinction in medicine. *Medicine studies*, *1*(2), 167-181.

Kullo, I. J., Lewis, C. M., Inouye, M., et al. (2022). Polygenic scores in biomedical research. *Nature Reviews Genetics*, 1-9.

Lee, J. J., Wedow, R., Okbay, A., et al. (2018). Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. Nature genetics, 50(8), 1112-1121.

Levy, A. (2021). Idealization and abstraction: refining the distinction. *Synthese, 198*, 5855-5872.

Lewis, A. C., Green, R. C., & Vassy, J. L. (2021). Polygenic risk scores in the clinic: translating risk into action. *Human Genetics and Genomics Advances*, *2*(4).

Lewis, C. M., & Vassos, E. (2017). Prospects for using risk scores in polygenic medicine. *Genome medicine*, *9*, 1-3.

Lewontin, R. C. (1974). Annotation: the analysis of variance and the analysis of causes. *American journal of human genetics*, *26*(3), 400-411.

Love, A. C., & Nathan, M. J. (2015). The idealization of causation in mechanistic explanation. *Philosophy of Science*, *82*(5), 761-774.

Lu, Q., & Bourrat, P. (2022). On the causal interpretation of heritability from a structural causal modeling perspective. *Studies in History and Philosophy of Science*, *94*, 87-98.

Madole, J. W., & Harden, K. P. (2022). Building causal knowledge in behavior genetics. *Behavioral and Brain Sciences*, 1-76.

Marchini, J., Cardon, L. R., Phillips, M. S., & Donnelly, P. (2004). The effects of human population structure on large genetic association studies. *Nature genetics*, *36*(5), 512-517.

Marees, A. T., de Kluiver, H., Stringer, S., et al. (2018). A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *International journal of methods in psychiatric research*, *27*(2), e1608.

Martin, A. R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B. M., & Daly, M. J. (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature genetics*, *51*(4), 584-591.

Mather, K. (1941). Variation and selection of polygenic characters. *Journal of Genet*ics, *41*(1), 159-193.

Mather, K. (1943). Polygenic Inheritance and Natural Selection. *Biological Reviews, 18*(1), 32-64.

Matthews, L. J. (2022). Half a century later and we're back where we started: How the problem of locality turned in to the problem of portability. *Studies in History and Philosophy of Science Part A*, *91*, 1-9.

Matthews, L. J., & Turkheimer, E. (2021). Across the great divide: pluralism and the hunt for missing heritability. Synthese, 198(3), 2297-2311.

Matthews, L. J., & Turkheimer, E. (2022). Three legs of the missing heritability problem. *Studies in History and Philosophy of Science*, *93*, 183-191.

McClellan, J., & King, M. C. (2010). Genetic heterogeneity in human disease. *Cell*, *141*(2), 210-217.

Meehl, P. E. (1992). Factors and taxa, traits and types, differences of degree and differences in kind. *Journal of personality*, *60*(1), 117-174.

Meehl, P. E. (1999). Clarifications about taxometric method. *Applied and Preventive Psychology*, *8*(3), 165-174.

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British journal of Psychology*, *88*(3), 355-383.

Michell, J. (2012). Alfred Binet and the concept of heterogeneous orders. *Frontiers in psychology*, *3*, 261.

Mostafavi, H., Harpak, A., Agarwal, I., et al. (2020). Variable prediction accuracy of polygenic scores within an ancestry group. *Elife*, *9*, e48376.

Nagpal, S., Gibson, G., & Marigorta, U. M. (2018). Pervasive modulation of obesity risk by the environment and genomic background. *Genes*, *9*(8), 411.

Nature Editorial (2022). The alarming rise of complex genetic testing in human embryo selection. Nature 603, 549-550.

Nelson, R. M., Pettersson, M. E., & Carlborg, O. (2013). A century after Fisher: time for a new paradigm in quantitative genetics. *Trends in Genetics*, *29*(12), 669-676.

NHGRI Online Resource (National Human Genome Research Institute). Polygenic risk scores. https://www.genome.gov/Health/Genomics-and-Medicine/Polygenic-risk-scores

NIH National Library of Medicine (2022). What are single nucleotide polymorphisms (SNPs)? https://medlineplus.gov/genetics/understanding/genomicresearch/snp/#:~:text=Single%20nucleotide%20polymorphisms%2C%20frequently%20called,building%20block%2C%20called%20a%20nucleotide

Oftedal, G. (2022). Proportionality of single nucleotide causation. *Studies in History and Philosophy of Science*, *93*, 215-222.

Owen, M. J., Craddock, N., & Jablensky, A. (2007). The genetic deconstruction of psychosis. *Schizophrenia bulletin*, *33*(4), 905-911.

Palk, A. C., Dalvie, S., De Vries, J., et al. (2019). Potential use of clinical polygenic risk scores in psychiatry–ethical implications and communicating high polygenic risk. *Philosophy, Ethics, and Humanities in Medicine*, *14*(1), 1-12.

Panse, V. G. (1940). A statistical study of quantitative inheritance. *Annals of Eugenics*, *10*, 76-105.

Schizophrenia Working Group (2014). Biological insights from 108 schizophrenia-associated genetic loci. *Nature, 511*(7510), 421-427.

Pedersen, L. H. (2017). The risks associated with prenatal antidepressant exposure: time for a precision medicine approach. *Expert opinion on drug safety*, *16*(8), 915-921.

Phillips, P. C. (2008). Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. *Nature Reviews Genetics*, *9*(11), 855-867.

Piccinini, G., & Craver, C. (2011). Integrating psychology and neuroscience: Functional analyses as mechanism sketches. *Synthese*, *183*(3), 283-311.

Pierce, B. A. (2017). *Genetics: A conceptual approach, Sixth Edition*. New York: Freeman & Company.

Plomin, R. (2019). *Blueprint, with a new afterword: How DNA makes us who we are*. Mit Press.

Plomin, R., DeFries, J. C., Knopik, V. S., & Neiderhiser, J. N. (2013). *Behavioral Genetics. Sixth Edition*. New York: Worth Publishers.

Plomin, R., Haworth, C. M. A., & Davis, O. S. P. (2009). Common disorders are quantitative traits. *Nature Reviews Genetics, 10*, 872-878.

Plomin, R., & von Stumm, S. (2022). Polygenic scores: Prediction versus explanation. Molecular psychiatry, 27(1), 49–52.

Portides, D. (2021). Thinking about idealization, abstraction and scientific models: an introduction. *Synthese, 198*(Suppl 24), 5849-5853.

Potochnik, A. (2020). Idealization and many aims. *Philosophy of science*, *87*(5), 933-943.

Purcell, S. (2013). Appendix. Statistical Methods in Behavioral Genetics. In R. Plomin, J. C. DeFries, V. S. Knopkin & J. N. Neiderhiser, *Behavioral Genetics. Sixth Edition* (pp. 357-411). New York: Worth Publishers.

Richardson, K., & Jones, M. C. (2019). Why genome-wide associations with cognitive ability measures are probably spurious. *New Ideas in Psychology*, *55*, 35-41.

Riglin, L., Collishaw, S., Thapar, A. K., et al. (2016). Association of genetic risk variants with attention-deficit/hyperactivity disorder trajectories in the general population. *JAMA psychiatry*, *73*(12), 1285-1292.

Robinson, E. B., St Pourcain, B., Anttila, V., et al. (2016). Genetic risk for autism spectrum disorders and neuropsychiatric variation in the general population. *Nature genetics*, *48*(5), 552-555.

Rutter, M. (2006). Genes and Behavior: Nature-Nurture Interplay Explained. Oxford: Blackwell.

Sarkar, S. (1998). *Genetics and reductionism*. Cambridge University Press.

Sauce, B., & Matzel, L. D. (2018). The paradox of intelligence: Heritability and malleability coexist in hidden gene-environment interplay. *Psychological bulletin*, *144*(1), 26.

Schaffner, K. F. (2016). *Behaving: What's genetic, what's not, and why should we care?*. Oxford University Press.

Scriver, C. R. (2007). The PAH gene, phenylketonuria, and a paradigm shift. *Human mutation*, *28*(9), 831-845.

Serpico, D. (2018). What kind of kind is intelligence? *Philosophical Psychology*, *31*(2), 232-252.

Serpico, D. (2020). Beyond quantitative and qualitative traits: three telling cases in the life sciences. *Biology & Philosophy*, *35*(3), 1-26.

Serpico, D., & Borghini, A. (2021). From obesity to energy metabolism: ontological perspectives on the metrics of human bodies. *Topoi*, *40*(3), 577-586.

Serpico, D., & Petrolini, V. (2023). Crossing the Threshold: An Epigenetic Alternative to Dimensional Accounts of Mental Disorders. *The British Journal for the Philosophy of Science.*

Serpico, D., Lynch, K. E., & Porter, T. M. (2023). New Historical and Philosophical Perspectives on Quantitative Genetics. *Studies in History and Philosophy of Science*.

Strevens, M. (2021). Permissible idealizations for the purpose of prediction. *Studies in History and Philosophy of Science Part A*, *85*, 92-100.

Tabery, J. (2014). *Beyond versus: The struggle to understand the interaction of nature and nurture*. mit Press.

Trejo, S., & Domingue, B. W. (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.

Turkheimer, E. (2011). Still missing. *Research in Human Development*, *8*(3-4), 227-241.

Turkheimer, E. (2022). Peter Visscher on the Genomics of Complex Human Traits. http://turkheimer.com/peter-visscher-on-the-genomics-of-complex-human-traits

Turley, P., Meyer, M. N., Wang, N., et al. (2021). Problems with using polygenic scores to select embryos. *New England Journal of Medicine*, *385*(1), 78-86.

Uffelmann, E., Huang, Q. Q., Munung, N. S., et al. (2021). Genome-wide association studies. *Nature Reviews Methods Primers*, *1*(1), 1-21.

Van Fraassen, B. C. (1980). *The scientific image*. Oxford University Press.

Wahlsten, D. (1994). The intelligence of heritability. *Canadian Psychology/Psychologie canadienne*, *35*(3), 244.

Wang, Y., Tsuo, K., Kanai, M., et al. (2022). Challenges and Opportunities for Developing More Generalizable Polygenic Risk Scores. *Annual Review of Biomedical Data Science*, *5*.

Webber, C. (2017). Epistasis in neuropsychiatric disorders. *Trends in Genetics*, *33*(4), 256-265.

Wehby, G. L. (2018). Genetic predisposition to obesity widens and skews the BMI distribution. *Obesity research & clinical practice*, *12*(4), 401-402.

Weisberg, M. (2012). *Simulation and similarity: Using models to understand the world*. Oxford University Press.

Woodward, J., & Kendler, K. (2023). Polygene Risk Scores: A Philosophical Exploration. *Philosophy of Medicine*, *4*(1), 1-21.

Zaidi, A. A., & Mathieson, I. (2020). Demographic history mediates the effect of stratification on polygenic scores. *Elife*, *9*, e61548.