The Scope and Limits of Reasoning Practices in Behavioral Genetics

A Dissertation

SUBMITTED TO THE

FACULTY OF THE

UNIVERSITY OF MINNESOTA

BY

Lauren Hunter Wilson

IN PARTIAL FULFILLMENT OF THE

REQUIREMENTS FOR THE DEGREE OF DOCTOR

OF PHILOSOPHY

Alan C. Love

2025

Acknowledgments

I have had a fantastic experience in graduate school. This is in no small part due to my incredible advisor, Alan Love. I am continually amazed by Alan's ability to juggle so many different irons in the fire and yet still make time to provide rigorous feedback to both myself and other graduate students. Yet one of the things I admire most about Alan is his skill in crafting an intellectual community. He has a wonderful ability to bring together people with wide ranges of disciplinary backgrounds – from graduate students to emeritus professors – and get them to connect and really talk with one another. So much of what makes these meetings successful happens is due to all the work Alan puts in beforehand. To see Alan in action crafting these meetings is inspiring.

One intellectual community that Alan has beautifully crafted is his advising group. Advising group, or "Love Lab" (complete with lava lamp) as we affectionately referred to it, was a constant source of support in the form of constructive feedback. I am so grateful to all members of Love Lab. Max Dresow, Nathan Lakey, and Yoshinari Yoshida were founding members of my Love Lab experience. I could not have asked for better people to shape my early years in graduate school. I did and still do look up to them so much. Max has the beautiful ability to tell rich philosophical stories. I aspire to be able to write like Max. Nathan Lackey helped me understand my own writing process better and taught me to be confident in my abilities, even as I was still learning. Yoshi's attention to detail and creativity in both his philosophical work but also his artistry is amazing. He approaches things with such curiosity. The early years of Love Lab also included visiting student Kelle Dien and Postdoc Amanda Corris. I learned so much from Kelle about ethology and the study of animal behavior. I am so grateful to have worked with Amanda on several projects and have learned so much. I am especially grateful to the postdocs in Love Lab: Correy Dethier, JP Gamboa, and Dana Matthiessen. They provided multiple rounds of feedback on each chapter of this dissertation which was essential for me in articulating exactly what I wanted to accomplish. They were also instrumental in helping my thoughts get to the pages of this document.

Another intellectual community I am incredibly grateful for is the Biology Interest Group (BIG). BIG was always my favorite part of the week. I always left inspired by our conversations and in awe of the way everyone from undergraduates to emeritus professors could come together to discuss (and debate) biology. I am grateful especially to Anya Auerbach, Mark Borello, Ron Okagaki, Mike Travisano, and Marlene Zuk for their probing questions and keen insight.

I am incredibly grateful to my committee members: Carl Elliott, Bennett McNulty, Ruth Shaw, and Valerie Tiberius Valerie Tiberius taught me about the many ways philosophers have understood happiness and wellbeing, ideas which operate in the background of my work as I try to lay out the ethics of the sciences I engage with. I deeply admire Ruth and

all the work she has done to push back against themes of genetic determinism both with her students and also in her own scientific work. Bennett has the gift of asking carefully crafted questions and the ability to see the connections between complex and seemingly desperate ideas. In addition to being an amazing committee member, Bennett is an incredible teacher. I am continually inspired by his ability to create a thriving learning environment for students, one which prioritizes students' intellectual curiosity and supports them in finding connection to philosophy in novel ways. Carl embodies the best of what bioethics is meant to be and do in the world. I am inspired by his dedication to bring to light the horrors of scientific research and fight to bring justice to those most vulnerable in society. I am incredibly grateful to have learned bioethics from someone who not only teaches bioethics but lives by the same values. I am not only grateful for Carl's mentorship in bioethics but also his kindness.

I would also like to express appreciation to the members of the Department of Philosophy. Here, I was able to broaden my philosophical knowledge and learn to engage with many branches of philosophy. Thank you to Peter Hanks, Valerie Tiberius, Michelle Mason Bizri, Jessica Gordon-Roth, Tamara Fakhoury, Roy Cook, and Cat Saint-Croix. I am especially grateful to Samuel Fletcher. He taught my first formal methods class which was a struggle. However, he recognized I was dedicated to improving my understanding and offered me the opportunity to help him improve the course, especially for students like me with no formal background in mathematical methods in philosophy. Samuel Fletcher also encouraged me to apply for the Yudof Fellowship which I received in my third year. I am especially grateful to fellow graduate student Ray Pederson who was instrumental in instigating community and sharing the joy of thinking critically about teaching

I would not have gone to graduate school without the help of the amazing faculty and Whitman College, especially: Nancy Forsthoefel, Dan Vernon, Mitch Clearfield, Michelle Jenkins, Rebeccaa Hanrahan, and Patrick Frierson. Nancy and Dan gave me my first taste of functional genetics research in their plant lab (affectionately called "team weed"). My experience working in this lab deeply informs my view of science and my commitment to scientific practice. Working there also shaped my views on the importance of working with model organisms, like *A. Thaliana.* Dan is also the person who taught much of my undergraduate genetics and molecular biology classes and I still hear SNP's in his voice whenever I read it (for better or for worse, it's a very silly voice). Mitch Clearfield allowed me to be a writing fellow for a truly unique course that brought together 16 traditional Whitman students and 16 students from the Washington State Penitentiary to discuss complex ethical cases. I learned so much from Mitch about how to build community across differences and get students to connect with one another. I implemented many of the things I learned from Mitch in my classes today. Michelle Jenkins taught my first collegiate philosophy class: ethics and bioethics. I didn't realize until much later how powerfully she shaped my understanding of philosophy, especially the importance of normalizing and uplifting previously marginalized subfields like feminist philosophy. She taught feminist care ethics right alongside deontology and forever legitimized feminist ethics for me. She also approached her students with kindness in a way I hope to emulate. Rebecca Hanrahan

has one of the sharpest wits I have encountered. She taught me how to break down an argument and build a better one in its place. I am so grateful for Rebecca's continued mentorship. She continues to be the person I reach out to share both the joy and also lows of academia. In some sense, Patrick Frierson is the reason I went to the University of Minnesota and the reason I study Philosophy of Science. Patrick invited me to do an independent study my sophomore year on philosophy of biology (he actually used Alan's Syllabus to do this). It was the hardest course I have ever taken but also made me fall in love with philosophy of biology. Patrick also coached Ethics Bowl which allowed me to hone my public speaking skills and build community with others interested in philosophy. I am also grateful for Partick's support in applying to graduate school. He made it clear that I had the skills to bridge both ethics and philosophy of science in a way that could be an asset in the future. I also want to acknowledge Gillian Grey, a philosophy major at Whitman and member of Ethics Bowl a few years older than me. Gillian was one of the first people from our college to go to graduate school for analytic philosophy and help me both apply to study at Oxford as a junior year abroad student and later apply to graduate school. Gillian also showed me you can be an incredible philosopher while also being deeply kind.

I also want to acknowledge four teachers who helped in my early learning: Matt Wright, Amy Fujimoto, Sunita Pailoor, and Sharron Dunn. Matt Wright taught my first philosophy class in high school through the Matteo Richi Program at Seattle University. He taught me the value of asking questions in seminars instead of pontificating. Amy and Sunita were my elementary school teachers and helped me through many difficulties while keeping my confidence intact. I appreciate all of the extra support they provided me. Sharron taught me how to write. She was the first person to make my papers bleed red and taught me how to take constructive feedback and use it to improve my work. She also provided guidance on my 9th grade yearlong passion project which was about genetic engineering in plants (I later got the genetically engineered plant in Dan's Lab at Whitman) and its ethics. I still have the little stuffed cat Sharron gave me when I graduated from Woodinville Montessori on my desk at home where much of this dissertation was written.

On a slightly silly note, I want to thank Hugo, my golden retriever. I got him one year into my PhD program and he has been one of my greatest supporters, staying up laying at my feet through many late night writing sessions. He reminds me that there is more to life than work and that even the worst writing day can be made better by going on a walk.

I am grateful to my partner, Cameron Conner. He has seen this whole process from choosing to apply to PhD programs in Philosophy, through course work, thesis proposal, and now the defense. We have lived both together and apart but through it all he has provided so much support along this journey. He too is incredibly driven and passionate about his work and reminds me how amazing it is that I get to do philosophy, something I love, every day.

Last but certainly not least I am incredibly grateful for my parents, Jeff and Dena Wilson, and my brother Jeffrey Wilson. They have always believed in me. They taught me the importance of education and love of knowledge. They supported my curiosity and let me learn and grow at my own pace. They have always encouraged my love of reading, philosophy, and the natural world. I truly cannot put into words all they have done for me and how much they mean to me. I could not have done this without them. They mean the world to me.

Dedication

To my parents, for always believing in me.

Abstract

Although philosophers have engaged with human behavioral genetics and the epistemic role model organisms play in research, little attention has been paid to human behavioral genetics research performed in model organisms. My dissertation, **"**The Scope and Limits of Reasoning Practices in Behavioral Genetics"**,** investigates the conceptual and ethical issues that arise when viewing human behavioral genetics in light of behavioral genetics performed in non-human organisms. **Chapter 1** motivates the project, arguing that if we accept that non-human model organisms can provide experimental traction for research in human contexts then the same strategy ought to apply to behavioral genetics. Accepting this claim opens new domains of inquiry about behavioral genetics, both practically and philosophically. **Chapter 2** examines what counts as a behavioral trait for the purpose of scientific investigation. I develop a cluster of criteria that are neither necessary nor sufficient but provide a reasonable framework for the scientific investigation of behavioral traits. As a result, a basis for normative evaluations emerges that rules some things in (e.g., learning and memory in fruit flies) and other things out (e.g., educational attainment in humans), with attendant consequences for different social and behavioral sciences. In **chapter 3**, I develop the concept of methodological shortfall from other existing explanations of science behaving counterintuitively. Methodological shortfall occurs when research communities concentrate their research programs around research questions, but their available methods are constrained in their ability to answer these questions unambiguously. In other words, a research community's reach exceeds its grasp. I use contemporary behavioral genetics as a case study to demonstrate how the concept of

methodological shortfall can help us make sense of ongoing complaints about behavioral genetics research undertaken on humans. Analyzing these issues through the lens of methodological shortfall reveals two paths forward for human behavioral genetics research: (1) continuing to use observational studies like Genome Wide Association Studies (GWAS) and Twin Studies, pursuing incremental improvements such as increased sample sizes, or (2) adopting experimental methodologies that can only be used in model organisms, which means approaching their object of study in an indirect way. After characterizing these two paths, I journey down each one to highlight how that choice impacts epistemic progress. Some researchers in behavioral genetics have chosen to work with model organisms to overcome the methodological barriers inherent in observational studies. I show how their work has generated epistemic progress on central research questions that have not been achieved by those working with human populations. I close with some observations about how the concept of methodological shortfall can contribute to broader discussions within philosophy of science. **Chapter 4** directly addresses the question of the limits of reasoning practices in behavioral genetics, especially regarding claims that it can help create a more just society. I concentrate my argument on the epistemic merits of the proposed use of behavioral genetics in the context of education. I leverage analyses from my previous chapters to identify two necessary conditions that behavioral genetics would need to meet before the bioethics community could reevaluate whether it should be used to address social inequality. I then consider if a future field of behavioral genetics, one that has implemented these conditions, would have anything to offer. I leave open the possibility that the field may one day have something to contribute

toward efforts to mitigate social inequality but highlight the potential harms of discrimination that are currently left unaddressed and may be exacerbated if behavioral genetics is deployed in the interim.

Contents

Contents

List of Figures

Chapter 1: Revisiting Behavioral Genetics: Model Organisms, Methodological Disputes

1.1 Introduction

Behavioral Genetics (BG) is a branch of quantitative genetics that focuses on differences between individuals and asks whether these differences are better accounted for by genetic or environmental differences (or some combination thereof). The field has faced persistent criticisms from different quarters, including from historians, philosophers, and other scientists. Within philosophy of science, however, this interaction has primarily concerned a narrow set of conceptual issues and controversies involving research performed in human contexts, such as the heritability of IQ and its consequences for race and intelligence (Lewontin 1974; Downes and Matthews 2020). Yet human populations do not serve as the only research subjects for BG. There is a long history of using non-human model organisms in BG research (Panofsky 2014), which is largely overlooked by philosophers. For example, even though Kenneth Schaffner's comprehensive treatment provides a detailed overview of many philosophical issues in BG and touches on research in non-human model organisms (Schaffner 2016), it still spends most of its time on human-focused research. Very little philosophical work has been directed towards epistemic issues related to BG research performed in non-human organisms.

In my dissertation, I identify a missed opportunity in both the scientific and philosophical literatures. By focusing their conceptual critiques on studies performed in

human populations, scholars have ignored the contributions and significance of studies undertaken in non-human model organisms. In this they unwittingly follow a pattern of reasoning that is also evident among human BG researchers, who similarly ignore these forms of research. The pattern is especially interesting given the epistemic role that studies of non-human model organisms play in other human-oriented investigations, like the study of conserved molecular mechanisms in biomedicine and neurobiology. The glaring methodological discrepancy between what is observed in human BG and what is observed elsewhere in life science investigation constitutes the gateway into my philosophical analysis.

Each chapter of the dissertation addresses one facet of expanding the philosophical conversation about BG to incorporate non-human model organisms. Chapter 1 sets the stage with my central argument: If we accept that non-human model organisms can provide experimental traction for research in human contexts (e.g., physiological traits in biomedicine), then the same reasoning strategy ought to apply to BG. I show that current scientific and philosophical literature on BG has not engaged with work in non-human model organisms and then highlight several consequences of this situation. Chapter 2 explores what counts as a behavioral trait for the purpose of BG research. To this end, I offer a cluster of criteria that are neither necessary nor sufficient but provide a reasonable strategy for the scientific investigation of behavioral traits. Adopting these criteria provides a basis for normative evaluations that rule some things in (e.g., learning and memory in fruit flies) and other things out (e.g., educational attainment in humans), with attendant consequences for different social and behavioral sciences. Chapter 3 explores seemingly

counterintuitive pattern in human BG—the continued use of specific empirical methods despite their insufficiency to discriminate between alternative hypotheses. I refer to this phenomenon as methodological shortfall because the human BG community's reach exceeds its grasp. Researchers that encounter this predicament face two options: (1) work with or improve available methods in an attempt to overcome the existing shortfall (including elaborating on and tinkering with existing methods) or (2) adopt different methodological approaches. This second option is often more time-intensive and can involve learning new techniques and redirecting existing resources. Chapter 4 examines the limits of reasoning practices in BG, especially in terms of how they can or cannot inform policy or individual decisions. It explicitly focuses on claims made by researchers that "understanding the role of luck in our lives—including the role of genetic luck—is vital to the egalitarian project" and that BG provides the tools to do so (Harden 2021a, 248).

This introductory chapter initiates the project with my core argument. If we accept that non-human model organisms can provide relevant evidence to discriminate between competing hypotheses in other areas of human research (e.g., physiological traits in biomedicine), then the same reasoning strategy ought to apply to BG. If the antecedent is accepted then it suggests several conceptual consequences, including the need to increase the range of considerations for engagement with BG research. Section 1.2 presents the current state of philosophical work on BG, highlighting how philosophers primarily focus on issues in human research. Section 1.3 explores the role of non-human model organisms in other fields of biology and shows that, although BG has many of the same features and

limitations that motivate the use of non-human model organisms, it does not operate like these other fields. Section 1.4 compares researchers working on human and non-human models, focusing on differences in disciplinary backgrounds, epistemic commitments, aims, and methods. Finally, Section 1.5 suggests what may be gained, both practically and philosophically, from engaging with BG research conducted in non-human organisms and offers potential directions for future research.

1.2 Existing Philosophical Work on Behavioral Genetics

There is a flourishing area of philosophy of science and bioethics concerned with the genetic study of human traits that has long been of interest to philosophers, especially since the completion of the Human Genome Project. This work includes but is not limited to debates about nature versus nurture. For example, Helen Longino and Evelyn Fox Keller (Longino 2013; Keller 2009; 2017; 2014) have analyzed what can be known through different kinds of empirical investigation involving different assumptions and methodologies in BG. However, there are three important characteristics of current philosophical engagement with the field of BG and its central concepts such as heritability.

First, although many scholars have scrutinized the successes and shortcomings of BG studies of human traits (Tabery 2014; Griffiths and Tabery 2008; Tabery 2007; 2015a; 2009a; Matthews 2022; Tabery 2009b; Plaisance 2006; Plaisance and Reydon 2012; Plaisance, Reydon, and Elgin 2012; Burt, Plaisance, and Hambrick 2019), these philosophical discussions have limited the scope of their inquiry to knowledge gained through research done exclusively in human populations. This may be explained in part by

the increased attention given to issues surrounding race and IQ in BG (Gould 1981, Paul 1998;Segerstråle 2000). However, it becomes increasingly difficult to justify this limited scope once one identifies the main target of philosophical engagement with human BG.

Philosophical engagement with BG has been fixated on issues, especially limitations, surrounding the methods of investigation used in human BG research. Critics note that the available methods used by behavioral geneticists in humans do not generate the kind of knowledge desired (Coop and Przeworski 2022b; 2022a; Harden 2022; Riskin and Feldman 2021). Others have raised concerns about the inconsistency of different heritability estimates in humans yielded by different methods. One example is the discrepancy between estimates provided by twin studies and those utilizing Genome Wide Association Studies (GWAS), generally referred to as the problem of missing heritability (Maher 2008; Matthews and Turkheimer 2021; Plomin and von Stumm 2018; Zuk et al. 2012). Furthermore, researchers disagree on the partitioning of shared and non-shared environments, on the identification of relevant variables, and on whether predictions made based on one data set transfer or export to populations not included in the original data set (Matthews and Turkheimer 2022; Matthews 2022; Plaisance 2006).

Another key locus of philosophical engagement is that the results derived from methods used to study humans are confounded by the existence of population stratification, which refers to systematic differences in allele frequencies between groups due to different ancestry rather than causally bearing on trait differences. The main way to disrupt population stratification and limit this cofounding is to experimentally assign mates at

random among groups (i.e., choose who mates with whom). This practice, of course, would be highly unethical in the case of humans. It is noteworthy that non-human model organisms also exhibit population stratification. However, in non-human model organisms it is ethically permissible and practically feasible to intervene to disrupt population stratification to separate genuine causal relationships from spurious correlations.

To date there has been very little attention given to non-human model organisms in BG despite a long history of quantitative genetic research on behavior in organisms other than humans. This lapse is surprising in view of ongoing discussions about the use and misuse of model organisms in a variety of domains of life science, including molecular genetics, biomedicine, and neurobiology (Bier and McGinnis 2008; Ankeny and Leonelli 2011; Nelson 2018; Ankeny and Leonelli 2021). Philosophical engagement with practices surrounding non-human model organisms in other domains of biology is motivated by the shared recognition that the study of human populations is constrained by practical and ethical considerations, such as an inability to experimentally assign mating pairs. Furthermore, these fields share a commitment to studying model organisms because they can and have helped the field learn something about traits of interest, especially through experimental intervention. As a result, the methods used in non-human model organism research within areas like biomedicine are more powerful in their empirical reach than those available in studies working only with humans. In light of this general recognition that model organisms can teach us something pertinent about a human trait of interest, it is important to take stock of epistemic practices related to both human and non-human taxa

and the relationships between claims made about human and non-human animals (or even plants) in BG.

Several sociological treatments of BG make note of model organisms and their role in the founding of the field (Panofsky 2014). However, these do not situate their treatment of BG in the context of the epistemic practices of researchers working either on human or non-human model organisms. The focus is on what motivations and incentives drove the community rather than exploring how different epistemic commitments, like the role of experimental intervention or a subject's ability to control its environment, influenced the scientific practice of researchers. Kenneth Schaffner provides a detailed overview of many philosophical issues in the field and touches on BG research in non-human model organisms (Schaffner 2016). Notwithstanding the value of Schaffner's work, it represents a third common limiting feature of philosophical engagement with BG. When philosophical work does explore the usage of model organisms in BG, the analysis is restricted to a small subset of organisms (mainly *Caenorhabditis elegans*, a roundworm species found in soil) designed to closely mimic particular human behaviors. Furthermore, the discussion of empirical studies conducted in these organisms is leveraged to gain traction on debates surrounding issues of developmental systems theory, human pathology, and candidate gene methods (Shaffner 2016), without emphasis on the epistemic issues surrounding the use of model organisms in BG itself. Thus, while previous philosophical analyses have used the results of research in non-human model organisms to interrogate ongoing debates in other domains of biology, what we still lack is a philosophical treatment of epistemic issues surrounding BG research in non-human organisms.

Having taken stock of the current philosophical literature, three key features are salient: (1) The overwhelming majority of scholarship focuses on issues related to BG research in humans, *per se*. (2) Scholars focus on methods used in human populations and issues of population structure, but have not explored how these issues operate in non-human model organisms where population structure can be intervened upon. (3) Philosophical work on model organisms has been limited to engaging with a few taxa that have been developed to directly reflect human conditions with an eye to questions outside of BG itself (e.g., developmental systems theory). Given these three features of the relatively narrow philosophical engagement with BG to date, it is worth examining what we might expect a broader engagement between human and model organism researchers to look like in BG.

1.3 What Ought We Expect Behavioral Genetics to Do?

Philosophers have not engaged sufficiently with model organism research in BG due, in part, to a lack of engagement between BG researchers working in human and model organisms, respectively. One may be tempted to explain this discrepancy by claiming that there is simply nothing to be gleaned through this engagement. In short, some people may think that the genetics underlying trait variation in *Drosophila* (fruit flies) is not relevant to humans. However, this line of reasoning is highly tenuous. For example, model organisms allow BG researchers to utilize interventionist techniques to disrupt population stratification, a key confounding factor in attempts to interpret BG results in humans.

Historians and philosophers of science have thoroughly examined the ways in which specific model organisms have helped to give insight into the phenomena or

mechanisms of different taxa (Ankeny 2010; Ankeny and Leonelli 2021; Ankeny et al. 2014). Philosophical investigation of the use of model organisms in scientific practice is ongoing, but there is already a consensus that they are generally useful—we learn something from their use in scientific practice. The points of disagreement in this area concern *what* and *how much* we learn, especially given the variable attention we pay to research on specific model organisms in different situations. Research projects focused on model organisms play a large role in research aimed at understanding humans because there are substantial practical and ethical limitations on human research. Non-human model organisms accomplish this role as either exemplars or surrogates (Bolker 2009).

Exemplary model organisms refer to organisms that, when studied, are taken to be representatives of a large group of organisms. For example, researchers often see zebra fish as a model of vertebrates. Results regarding a mechanism or phenomenon in an exemplary model organisms are generalized to the group of which the organism is taken to be a member. Conventional model organisms in developmental biology are exemplary in that "The motivation for their study is not simply to understand how that particular animal develops, but to use it as an example of how all animals develop" (Slack 2006, 61). However, model organisms can also be used as substitutes for biological systems of interest. When organisms are used in this way the results from their study are applied to a specific target. For example, mice are used as a model for *Homo sapiens* (Nelson 2018). The inference in this case is from a proxy to a target instead of from an exemplar to a larger group.

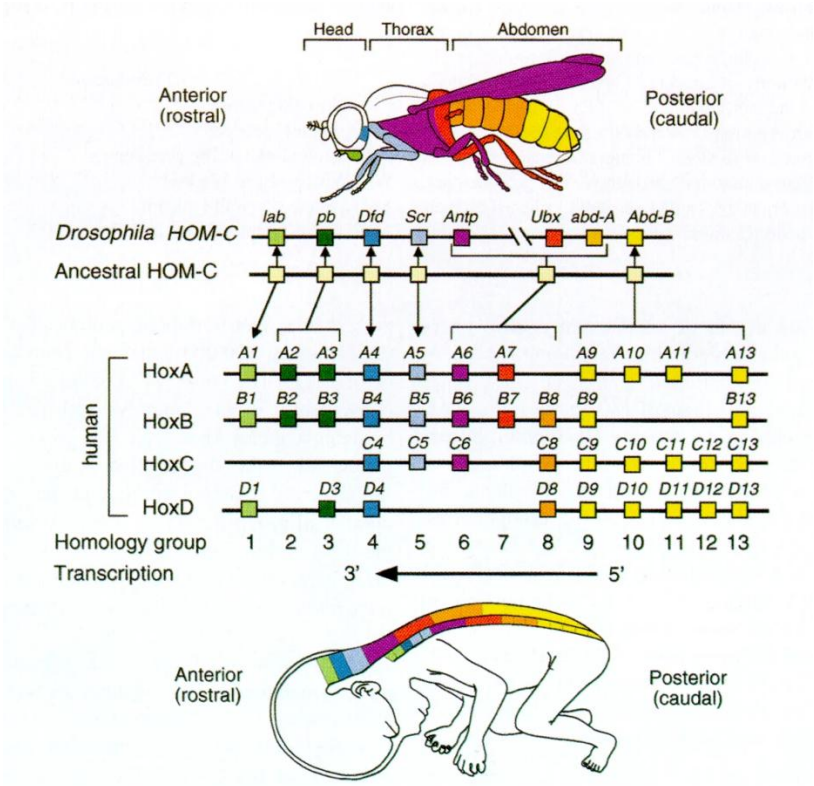

Figure 1: a schematic representation of the genomic organization and expression patterns of Fruit fly and Human Homeobox (HOX) Genes

Note that humans have four homeobox groups (A, B, C, and D) due to past rounds of genomic duplication in vertebrates. The representation also includes a hypothetical, ancestral homeotic complex with possible relationships to extant homeobox genes. Each gene is represented by a colored box (Mark, Rijli, and Chambon 1997 reproduced with permission from Springer Nature).

Using model organisms, even those that are evolutionarily distant, to learn about humans is neither controversial nor strange. Exemplary model organisms have played a pivotal role in developmental biology and specifically in the discovery of Homeobox (HOX) genes. Homeobox genes are a subset of genes that have been found to specify regions of the body plan in an embryo along the anterior (rostral/head)-posterior (caudal/tail) axis (Figure 1). HOX genes ensure that major structures (e.g., limbs) develop in the usual number at the

usual positions along the body. For example, HOX genes control development such that your arms develop towards the anterior of your body rather than near your pelvis.

Almost every developmental biology textbook contains a figure of a human embryo with colors matching a region of the body with a particular HOX gene that is activated there. However, there are clear ethical and practical limitations on the methods needed to experimentally manipulate human embryos to discover this relationship between HOX genes and development in humans. So how did researchers identify this connection between specific HOX genes and particular regions of bodily development in humans? Much of our knowledge comes from studies in *Drosophila* (fruit flies) that have explored the relationship between the arrangement of the HOX genes on the chromosome and their expression in the body (Gummalla et al. 2014; Lewis 1978). Fruit fly researchers relied on the ability to experimentally manipulate their model system by generating mutants and then mapping the physical malformations to corresponding differences at specific locations on the chromosome. Researchers found that these locations (subsequently identified as gene regions) are broadly conserved across species by identifying and comparing these regions across taxa, including humans. *Drosophila* served as an exemplary model in this situation.

Developmental biology serves as one of many areas in which non-human model organisms have been used to provide information about humans. Biologists and philosophers generally accept the idea that model organisms can provide experimental traction for research into questions about human biology. However, BG operates as an exception to this view. Scholarly articles on BG research performed in humans often have few or no citations of studies performed in model organisms, even for traits that may have

conserved mechanisms. For example, a *Science* paper using Genome Wide Association Studies (GWAS) to study same sex sexual behavior in humans contains only one citation of work performed in non-human model organisms, a mouse study on gonad development (Ganna et al. 2019). There were no citations of the numerous studies of sexual behavior in model organisms that utilize GWAS or experimental methods, which is surprising given the long history of these studies in organisms like *Drosophila*.

When human BG researchers do cite results from model organisms, they use this research to annotate the functions of regions of the genome found to be significant through GWAS performed on human genomes. Consider studies of educational attainment (EA) in humans. Lee et al. (2018) published a GWAS of educational attainment based on a sample size that was among the largest to date. The results of the study generated a polygenic score (an aggregation of estimated effects from across the genome) that explains 11% of the variance in EA. (The variance is a measure of the dispersion of the distribution of variation.) The authors highlight genes of interest related to brain development associated with some of these effects, especially those implicated in neuron-to-neuron communication, as worthy of further investigation. Their analysis relied on existing research from non-human model organisms that described the function of these genes (e.g., their relation to ion channels, metabolic pathways, and synaptic plasticity) and when the genes were expressed in development (both pre- and postnatally). Selecting out these genes for prioritization from among thousands contained in the polygenic score depends on understanding their function and how this may fit within a (currently poorly defined)

mechanism of educational attainment. Ironically, this understanding of gene function derives from both studies in humans and model organisms, especially mammals like mice.

Yet, within the authors' broader discussion of educational attainment, model organism BG literature is cited, even for traits which may play a role in educational attainment, such as learning and memory. This is even more striking considering the long history of BG research on learning and memory in fruit flies. Compounding the strangeness of this citation practice is that BG researchers who work on model organisms point to human application as a potential implication of their results. For example, researchers interested in identifying genes that contribute to natural variation in learning and memory explicitly note the limitations of human studies to motivate their research to "understand how differences in these cognitive traits evolve among populations and species".

> So far, the only polymorphic genes thought to contribute to natural variation in learning performance, in any species, have been recently identified through polymorphism-association studies in humans. However, given the obvious constraints on human research, it will be difficult to study the evolutionary forces acting on these allelic variants and maintaining the polymorphisms. More insights into those forces could be gained from studying genes contributing to variation in learning performance in natural populations of model organisms, such as *Drosophila* (Mery et al. 2007, 13051).

Furthermore, the authors highlight experimental work performed in mammalian models, including mice, to further draw out the implications of their study on fruit flies. Thus, while human BG researchers limit their discussion to the human realm, researchers using model organisms cite studies from across diverse taxa, including using an interest in human populations to motivate their research.

If we accept that non-human model organisms can provide experimental traction for research in human contexts, then the same strategy ought to apply to BG research. In

this section, I have demonstrated that biologists and philosophers generally accept that model organisms can provide insight into other taxa, and specifically humans. The main debates address what exactly we learn and how much insight we can glean. However, I have also shown that human BG researchers do not substantively engage with non-human model organism research even though model organism researchers motivate their inquiry by showing its relevance to humans. There are good reasons why BG researchers should engage with model organisms research, but this does not currently occur. Why does this discrepancy exist?

1.4 Comprehending the Contours of Discrepancy

Previous scholarly work that interrogated the gap between human and model organism research in BG focused on understanding the field from a sociohistorical perspective. This approach relied on identifying key individuals and historical events, examining private and public correspondence in the discipline, and taking stock of high impact publications. In what follows, I show that the lack of engagement between human BG and model organism BG is not well explained by an historical lack of collaboration. In fact, the opposite is the case; researchers using non-human models played a critical role in the founding of the field; human and model organism BG researchers worked synergistically for many years, leaning on each other's strengths and compensating for acknowledged shortcomings. The sociohistorical dimensions of the gap between human and model organisms research does not provide a complete explanation of the discrepancy.

An important component of understanding the divide between these communities is a description of differences in their epistemic practices. By epistemic practices I refer to

the processes that communities engage in to generate scientific knowledge, including individual thought processes, social interaction, and material arrangements. Epistemic practices are deeply intertwined within sociohistorical dimensions of scientific communities, but they have often been backgrounded in work on BG. In what follows, I note differences in epistemic practices that are complementary to previous sociohistorical accounts of why these communities do not interact.

1.4.1 Shifting Disciplinary Backgrounds and Epistemic Commitments

Previous sociohistorical work on BG provides valuable components for an explanation of the discrepancy between human and non-human focused research. Sociohistorical studies have documented the crucial role played by model organism researchers in the early history of BG. The field, as it is recognized now, began in the mid twentieth century in response to the American eugenics movement. Although eugenic research programs made a home for themselves in the United States between 1900 and the mid-1940s, attitudes quickly shifted (at least publicly) after the atrocities perpetrated by Nazi scientists became known. This led to a change in the way researchers approached the genetic study of traits. The change manifested as a cohort of researchers across many different fields and in different populations coming to work together (Willoughby et al. 2022; Plomin 2023; Loehlin 2009).

Animal researchers played an important role in the early days when BG was coalescing as a field. Scott and John Fuller's 1965 study, *Genetics and the Social Behavior of the Dog,* was a defining publication. Furthermore, in this publication the authors considered omitting human research studies, although some were ultimately included (Scott and Fuller 1965; Panofsky 2014). This reflects some general tensions extant at this

point, such as the tension between the desire to say something about human behavior and the need for this research to be more rigorous, like what is found in animal experimental research. In the end, investigators prioritized relevance to human behavioral traits by focusing on aspects of behavioral physiology, such as nervous system functioning, memory span, and sense discrimination.

The main targets or interlocutors of this nascent field were behaviorist and environmentalist paradigms in the social sciences[1]. To challenge these paradigms, BG researchers focused their investigations on explaining individual variation and creating hybrid biosocial explanations of behavioral differences. This required the joint participation of scientists working in both human and non-human populations. Those working in human populations hoped that the participation of animal researchers "could be a model for and complement to human research" (Panofsky 2014, 44). Thus, at the outset and until the end of the 1970s, animal and human researchers participated in cultivating a robust transdisciplinary field where different kinds of research played critical roles. Those working in human populations gleaned insight from animal researchers on their study methods. Human researchers were inspired to work on behaviors closely related to physiology, rather than cultural traits, to avoid pitfalls of the eugenics movement and to capitalize on the measurements and experimental designs that animal researchers relied on. To do this, human researchers needed to see humans as another species, similar yet different from others, and thus placed humans within a comparative, evolutionary

[1] Behaviorists sought to explain behavior in humans and other animals in terms of conditioning or learning; environmentalists explored how different environments influence the variation in behavior

framework. Additionally, because human and non-human animal researchers were cultivating a collaborative field, this fostered a greater diversity of important research questions and results. In return, human BG researchers offered much needed motivation for funding this kind of research, especially via the National Institutes of Health, which animal researchers relied on.

This collaborative relationship came to an explosive end in the 1970s. It was upended by the publication and resulting conflict surrounding Arthur Jenson's 1969 publication *How Much Can We Boost IQ and Scholastic Achievement?* There Jenson argued that attempts to remediate educational inequalities would fail due to the outsized role that genetics plays in educational attainment. Especially charged was the connection made between the genetic propensities of different racial groups and their associated IQ scores. As many scholars have argued, this messy conflict reshaped the field of BG, driving a wedge between human researchers, who sought to defend their observational methods from what they perceived as an attack, and animal researchers who wanted to distance their work from such public scrutiny (Panofsky 2014; Tabery 2009b; Griffiths and Tabery 2008; Tabery 2015b).

Around this time, the field of BG also underwent demographic shifts. Data from a 1978 survey of the field, one of the only demographic surveys of BG to this day, is illuminating in this regard. The survey reported the number of courses completed across different areas of the sciences for both regular, emeritus, and associate members of the Behavioral Genetics Association (Plomin 1979). Additionally, regular and emeritus members were asked what formal training was recommended for a "well-trained behavioral

geneticist of the future". Regular and emeritus members were considered more established in the field and represented people who on average had completed their formal training more than a decade prior. Associate members were those who had not yet completed their doctorate. It is noteworthy that the data show that associate members closely matched the formal training recommended by more senior members. However, the training of younger members leans toward the behavioral sciences, with 16% of new members expected to have no formal training in biological sciences. This is compared to 9% having no behavioral sciences training (down from 23% in associate members). Thus, while training in biological sciences remained consistent, more young scholars were entering the field primarily with additional behavioral science training.

These sociohistorical insights about BG help us to identify insights into the field's changing epistemic commitments. Having fewer young scholars with a background in the biological sciences meant less of a commitment to a comparative, evolutionary framework. Instead of seeing humans as situated amongst other organismal taxa, humans were veiwed as more insulated and separate from other organisms (Panofsky 2014). Seeing humans in isolation also correlated with a shift to focus on traits that were perceived as more uniquely human. This meant the study of behaviors moved away from those close to physiology and towards behaviors typically seen as more cultural (e.g., religiosity). This shift in behavioral trait focus also allowed researchers to distance themselves from experimental methodologies developed in animals and instead utilize statistical tools that were particularly well suited for human studies.

This shift in epistemic commitments is still visible today and can be illustrated by comparing two recent studies: one performed on learning and memory in fruit flies, and the other on educational attainment in humans. BG is dedicated to exploring the genetic basis of complex behavioral trait differences; learning, memory, and educational attainment are considered complex behavioral traits. However, the way these traits are conceptualized and measured shows that they lie on different ends of a continuum between physiological and cultural traits.

Human BG research has a history of interrogating the genetic basis of differences in educational attainment (EA) (Lee et al. 2018; Willoughby et al. 2019; Rietveld et al. 2013; Harden 2021a; Herrnstein 1994; Plomin and von Stumm 2018; Branigan, McCallum, and Freese 2013). EA is usually operationalized as the number of years (or in some cases weeks or months) of formal schooling an individual has received or the highest degree they have acquired. This conception of the trait and its measurement relies on cultural markers: institutions, years of schooling, and formal degrees. Although EA may correlate or be affected by neurobiological (physiological) traits, such as ability to learn, memory, or attention, EA is also subject to cultural variables such as financial means, political stability (e.g., no obligation to perform government mandated military service), standardized test preparation, and family support. This dependence on cultural factors is seen more clearly when put into conversation with research into the basis of differences in learning and memory performed in fruit flies.

Model organism researchers must work with organisms that exhibits variations both in behaviors and in many different expressions of learning and memory. When researchers

study memory in *Drosophila melanogaster* (Anholt 2010; Anholt and Mackay 2004; Sokolowski 2001; Dubnau 2014), behavioral assays that involve avoiding different scents are critical to operationalizing the trait. Learning in *Drosophila* is measured as a physiological behavior, such as movement toward or away from a scent, as well as the presence or absence of genes or receptors that were previously identified as associated with learning. As will be explored in Chapter 2, the ways each community conceptualizes behaviors as traits differs so greatly that it is unclear whether a behavior examined in one community (human researchers) would be considered a trait by the other (model organism researchers). In other words, the two communities differ regarding the epistemic standards they apply to the conceptualization of behavioral

1.4.2 Aims of Investigation

Alongside these shifts in demographics and epistemic commitments came shifts in the aims of investigation for human and model organism researchers. Human investigators tend to focus on applied research. For instance, awareness of inequalities in society and a desire to identify their underlying causes motivates much of the research on EA. Some BG researchers, such as Kathryn Harden, have argued that understanding the genetic contribution to EA differences is a necessary condition for building a fair society (Harden 2021a). The investigation of the genetic basis of variation in this trait is motivated by policy concerns (e.g., specific interventions). Therefore, the aim of this research is to understand variation in behaviors like EA in terms of variation in environment and genetics to address societal inequalities. Using twin studies and GWAS, researchers produce heritability estimates (see below) and polygenic scores to obtain a kind of predictive measure for an

individual about their likelihood of developing a specific value of a behavioral trait (Willour et al. 2012; McGue et al. 2013; Anney et al. 2008; Bralten et al. 2021; Cesarini et al. 2009; Lee et al. 2018).

Although human researchers often tout potential interventions based on their research (e.g., specific policy recommendations), this applied aim has been strongly criticized. Critics argue that human BG is flawed on both epistemic and moral grounds; the methods used in these studies simply fail to yield the relevant kind of knowledge (Coop and Przeworski 2022b; 2022a; Riskin and Feldman 2021). Researchers point to the issue of "missing heritability," the gap between estimates of genetic influence derived from twin studies and those utilizing GWAS (Maher 2008; Matthews and Turkheimer 2021; Plomin and von Stumm 2018; Zuk et al. 2012). Furthermore, researchers disagree about how to accomplish the partitioning of shared and non-shared environments, the identification of relevant variables, and the transferability of prediction metrics based on one data set to populations not included in the data set (Matthews and Turkheimer 2022; Matthews 2022; Plaisance 2006).

While the intervention-oriented aims of human BG research breed controversy, BG studies of complex behavioral traits in other taxa do not face the same level of interrogation. This is particularly interesting since BG researchers working in non-human organisms, even those quite distantly related to humans (e.g., *D. melanogaster*), see their investigations as informing the genetic basis of complex trait differences *across* taxa. These researchers even draw connections between what is gleaned from their non-human models and the human domain. For instance, in the study on learning and memory in *D. melanogaster*

(Mery et al 2007), the authors explicitly highlight the similar function of the PKG protein in both *D. melanogaster* and mammals. Researchers note that mice deficient in PKG show deficits in cerebellum-dependent motor-learning tasks. By drawing this connection, researchers are attempting to engage in cross-taxa generalizations about the contribution of genetic variation to variation in complex behavioral traits.

One of the main motivations for studying the heritability of traits in humans is to assess how likely it is that either environmental or genetic interventions can succeed in altering expression of the trait of interest. In non-human research, the ascribed benefits are more varied. There is a general interest in understanding the contributions of genes and environment in the expression of behavioral traits. Furthermore, researchers aim to uncover shared mechanisms for producing behaviors across taxa and understand how these traits can be modified by evolutionary dynamics. Additionally, model organism research has practical applications for commercial animal and plant breeding (Bernardo 2008; Bergelson and Roux, 2010; Zhou et al. 2022; Yamamoto, Jallon, and Komatsu 1997).

1.4.3 Available Methods

There is a stark contrast between the methodologies available to human and model organism researchers. This is due to ethical constraints on human research, including the inability to control mate choice and proscription against doing so, limited ability to manipulate environmental conditions, practical limitations such as longer generation times, inability to collect pertinent amounts of environmental data, and lack of genotype

replicates. However, human researchers have attempted to navigate these constraints with limited resources, as can be seen in the study of EA.

In order to answer how much variation in EA is due to genetic variation in human populations, researchers have historically relied on twin studies (Downes and Matthews 2020). Twin studies operate by identifying a large number of pairs of monozygotic and dizygotic twins with some twins having both been raise in the same family and some raised apart.[2] Researchers then use survey data or longitudinal studies to track each twin's EA and some environmental details. When twins are reared in the same family, it is assumed that they either share much of the same environment or that they technically have a high degree of "shared environment." When twins are reared apart, they are considered to share much less of the same environment. The environment that differs between individual twins is considered the "non-shared" environment (Plaisance 2006). Researchers then use information about the shared environment, unshared environment, and genetic similarity (as represented in Falconer's equation), to calculate the relative genetic contribution to EA (Falconer 1996). Twin studies are used as a kind of "natural experiment" intended to represent a facsimile of interventionist experimental work, which is prohibited in humans. However, these kinds of studies are strictly correlational; they do not elucidate the causes of variation among particular individuals, and discovered associations only have applicability to a specific population.

[2] For these studies, researchers treat monozygotic twins as genetically identical, whereas dizygotic twins share roughly 50% of their genes. Thus, monozygotic twins are treated as more genetically similar then dizygotic twins.

Both human and model organism researchers can use molecular tools such as GWAS. These were first used in humans to study social traits in the early 2000s (Rietveld et al. 2013; Cesarini et al. 2009). GWAS allows researchers to identify variations in different physical regions of the genome that covary with behavioral traits of interest (Willour et al. 2012; McGue et al. 2013; Manchia et al. 2019; Anney et al. 2008; Ferguson 2010; Bralten et al. 2021; Brevik et al. 2016; Karlsson Linnér et al. 2019; Sokolowski 2001; Derringer et al. 2015; Dick et al. 2011; Sanders et al. 2017; 2021; Odintsova et al. 2019; Bulik-Sullivan et al. 2015; Bosker et al. 2011). Many researchers have argued that if we can identify variations in genetic sequences that are associated (correlated) with variations in a behavior, then we may be able to ascertain information about the genetic causes of trait variation. Although molecular tools like GWAS are available to all researchers, only model organism researchers can directly control the breeding of their experimental subjects and develop mutants through interventions. This is true of much of the research on the genetics of learning and memory in *D. melanogaster* (Markow 1987; Dubnau 2014; Sokolowski 2001; Yamamoto, Jallon, and Komatsu 1997; Mackay et al. 2005; Anholt and Mackay 2004; Anholt 2010). Mutant screens involve subjecting genetically similar flies to mutagens and then screening them for behavioral differences. Once behavioral mutants have been identified, researchers then attempt to uncover the molecular changes that give rise to the behavioral differences. Note that while researchers can perform observational studies in humans using methods like the behavioral assay, researchers cannot manipulate or generate novel genetic strains in humans, do not have access to hundreds of individuals that share almost identical genomes, nor can they (except in rare cases) access brain tissue

from the subjects in the behavioral assay (and even then, the sample sizes are normally very small). All these limitations in human research that do not apply to model organism research have historically fostered interaction between these approaches.

Scholars criticizing human BG research have appealed to studies performed in model organisms because they utilize these kinds of methods unavailable in humans. This can be seen in responses to Jenson's 1974 paper (Jensen 1974). For example, David Layzer's critique centered on the usage of heritability estimates (Layzer 1974). These are a correlational measure of how much variation of a trait in a population is attributable to variation in genetics. Layzer leverages empirical data from model organisms to argue that heritability estimates for the same trait differ drastically and systematically when ascertained through different methods. For example, he cites empirical work on heritability estimates for four traits in cattle (calf weight, milk yield, fat percentage, and body weight) that were calculated using four different methods: separated monozygotic twins, monozygotic and dizygotic twins, half-sister comparisons with environment randomized, and half-sisters at a field station. The estimates from separated monozygotic twins are close to unity, whereas estimates from half-sib correlations are low. There are possible explanations for this discrepancy, but deciding among them requires further information, information unavailable in the human research context due to ethical and practical limitations. Based on these kinds of empirical studies performed in non-human organisms, Layzer concludes that the heritability estimates leveraged in Jenson's work are likely grossly unreliable.

Richard Lewontin's critique focuses on the analysis of variance used by Jenson, and particularly the assumption of additivity (Lewontin 1974).[3] Lewontin reviews empirical research on norms of reaction (or patterns of phenotypic variation derived from placing the same genotype in different environments[4]), which were generated from empirical research on *Drosophila* by Dobzhansky and Spassky (Dobzhansky and Spassky 1944). The norm of reaction for larval variability at three different temperatures for fourth chromosome homozygotes, where the chromosomes have been sampled from a natural population, show a great deal of genotype-environment interaction with varying degrees of environmental sensitivity. Here it is critical to note that norms of reaction are impossible to investigate in humans because the same genotype cannot be observed in an array of different environments.[5] Although Lewontin acknowledges that little work had been done on norms of reaction for species in natural populations, he cites work on wild plant populations that shows considerable non-additivity. This pattern of reasoning is typical; appeal to empirical work performed in one group (wild plants) to address a limitation of work on another group (domesticated *Drosophila*). Lewontin argues that together with a knowledge of the relative frequencies of different human genotypes, a knowledge of norms of reaction could in principle predict the demographic and public health consequences of

[3] The assumption of additivity is that influences from genetic, shared environmental, and nonshared environmental are additive and separable.

[4] Geneotype environment interaction "is that part of the average deviation of individuals sharing the same environment and genotype that cannot be ascribed to the simple sum of the separate environmental and genotypic deviation" (Lewontin, 2006,521)

[5] While monozygotic twins separated at birth are the same genotype experiencing different environments to some degree, there is far less degree of control available to assess just how different the environments are from one another. Additionally, there is no replication, two of three individuals, whereas studies in *Drosophila* can have tens to hundreds of replications.

certain massive environmental changes. Analysis of variance, the kind used by Jenson, can do neither of these because its results are a unique function of the present distribution of environment and genotypes. Furthermore, the (legitimate) purpose of the analysis of variance in human genetics is to predict the rate at which selection may alter the genotypic composition of human populations and to reconstruct, in some cases, the past selective history of the species.

Both Layzer and Lewontin drew on research performed in model organisms to criticize methodology in human BG research. However, their interlocutors largely ignored this feature of their work. When their arguments did get traction with some scholars, it was because they utilized data from methods performed in model organisms.

From this brief discussion of BG methods used in human and model organisms, a few key observations can be made. First, researchers in human contexts rely on "natural experiments" that take advantage of instances where monozygotic and dizygotic twins have been raised in the same or different environments to make claims about the role of environmental or genetic variation in trait variation. These methods yield results that are associations, but they are often stretched to hint at causal claims. Both sets of researchers have access to certain molecular approaches like GWAS and can make associative claims. However, only model organism researchers have access to interventionist experimental techniques, specifically controlling who mates with whom, which can break up population stratification and thereby potentially lead to an understanding of how reliable heritability estimates really are and probe norms of reaction for behavioral traits across different environments.

1.5 The Practical Upshot

I have argued that if we accept the premise that model organisms can offer experimental traction for research relevant to human contexts, then the same strategy ought to apply to BG. Section 1.3 demonstrated that biologists in many different fields presume (with good reason) that studies of organisms other than humans offers experimental traction for research in human contexts. Furthermore, even non-human model organism researchers in BG justify their research by pitching its application to human contexts. However, there is almost no recognition amongst human BG researchers that research on organisms other than humans could lend insight into humans. I explain this discrepancy in terms of differences in disciplinary backgrounds, epistemic commitments, aims, and methods of these two research communities. Importantly, the fact that we can explain these discrepancies does not absolve human BG researchers from violating the norm of using the strategy. Once made explicit, the recognition that human BG researchers should draw on model organism research opens new questions for BG, both practically and philosophically. For example, other communities that straddle human and model organism systems share some of the features above and nevertheless engage in cross taxa discussion. Therefore, it is worth exploring the possibility that other conceptual issues are operating in BG, contributing to this divide.

One cluster of questions concerns the breadth of empirical inferences. Under what circumstances are scientists justified in making or utilizing generalizations across taxa? When is it permissible to engage in cross-taxa generalizations about seemingly similar traits (e.g., EA in humans versus learning and memory in *Drosophila*)? Answering these

questions involves exploring what mechanisms underlie traits across evolutionarily distant taxa, how these traits are measured (especially since most complex traits are measured using proxies), and how to make sense of different ways of operationalizing seemingly similar traits. It also involves tackling larger questions about how our intuitions shape scientific investigation, and how our perceptions of behavioral traits (across many dimensions of biology and culture) affect our intuitions about generalizing across taxa.

Another set of questions involves interrogating the relationship between available methods and the aims of researchers. Model organism researchers often have access to methods and experimental tools that human researchers will never be able to use. Many of the methodological issues highlighted by philosophers and critics of human BG are absent or minimized in non-human model organism research. Take, for example, the issue of gene and environment interactions. Is this issue present in research not subject to the same ethical and practical limitations (e.g., plants)? Given that researchers in plant genetics have access to methods used in human and non-human model organisms, one could attempt to cleanly distinguish the empirical reach of the methods. Additionally, one could document how particular inferences are justified (or not) considering available methods to inform when and how generalizations are warranted in contexts where intervention-based methods are unavailable (e.g., observational studies in human contexts).

Even if one rejects the conditional (i.e., 'If we accept that non-human model organisms can provide experimental traction for research in human contexts, then the same strategy ought to apply to BG'), new lines of inquiry are opened. If contemporary human behavioral geneticists reject this implication, it lays bare the burden of justifying their

rejection of it. One of their most compelling explanations for rejecting this claim is that the traits of interest are exceptional, informed by culture and particular to the human species, such that inferences from other taxa are unjustified. However, little evidence has been provided to support this claim even though it persists in the literature as an implicit and unchallenged assumption. Are the traits of interest to human BG researchers so distinct from those of other taxa that no form of generalization is possible? On the other hand, how should researchers formulate cross-taxa generalizations that are attuned to critical differences in both the trait and the taxon of study?

Finally, exploring the relationship between empirical findings in human and non-human taxa presents an opportunity to challenge existing paradigms in human BG. Those challenging the results of BG have traditionally leveled their critiques at the limitation of the methods used in humans, especially with respect to issues of population stratification. Unfortunately, this tendency for behavioral geneticists to overstate the reach of their empirical methods and certainty of results has become a kind of zombie, recurring across decades with no resolution in sight. How would engagement change if, instead of challenging the methods used in humans, one challenged human study results by using BG findings from model organisms?

These questions structure the content of the following chapters. The second chapter explores the conceptualization and measurement of behavioral traits across taxa. The third chapter analyzes methods used in human and non-human BG research, to assess how these methods shape or limit what we can know in this field. The final chapter applies this

analysis to establish the limits of reasoning practices in BG with respect to how it can or cannot inform policy recommendations or individual decisions.

Chapter 2: A Strategy for the Scientific Investigation of Behavioral Traits

2.1 Introduction

Kathryn Paige Harden reignited fierce debates in the social sciences with her book *The Genetic Lottery: Why DNA Matters for Social Equality* (Harden 2021). Harden, a psychologist and behavioral geneticist, uses genetic studies of behavioral traits to argue that progress on issues of social inequality cannot be made without attending to genetic differences. Critics like Marcus Feldman (a biologist with expertise in population genetics) and Jessica Riskin (a historian of science) object to the use of these genetic methods to study outcomes like educational attainment, social mobility, and personality that are relevant to social sciences. A key conceptual question underlies this debate: what counts as a *behavioral* trait for the purpose of scientific investigation?

To date, philosophers have largely neglected the issue of how behavioral traits are conceptualized (but see Brown 2014; Longino 2013), focusing instead on how to conceptualize *morphological* traits, especially in discussions of homology and evolutionary novelty (DiFrisco, Wagner, and Love 2023; DiFrisco, Love, and Wagner 2020). In what follows, I offer an answer to the question of what counts as a *behavioral* trait for the purpose of scientific investigation. The clause "for the purpose of scientific investigation" signals that the question is one of scientific methodology rather than

metaphysics (e.g., "What is a genuine or real trait?"). My answer takes the form of a strategy that gives guidelines for evaluating the conceptualization of behavioral traits, with special attention to the epistemic context of behavioral genetics (BG). I develop this strategy—the Analogical Contextual Strategy (ACS)—by drawing analogies to how researchers characterize morphological traits. For example, scientists often identify spatial boundaries ("joints") for morphological traits; temporal boundaries can be identified for some behaviors and can serve a similar role in individuation. In addition to temporal boundaries, other relevant factors include relationship to biological function, regularity of occurrence, and appearance at specified life history moments. Using these kinds of analogies, I offer a cluster of criteria that are neither necessary nor sufficient but provide a reasonable strategy for the conceptualization of behavioral traits in scientific inquiry. This strategy fulfills the desiderata of researchers studying behavior (and other traits), including their being useful for scientific work and having a general application.

To highlight the strengths of this strategy, I contrast it with two other approaches to behavioral trait conceptualization. The first is conventionalism, which supposes that behavioral traits (and traits more generally) are simply those that can be measured. The second approach is a mechanistic strategy that conceptualizes behavioral traits in terms of underlying shared molecular mechanisms, as is sometimes done with morphological traits. I argue that the first is too permissive (licensing almost anything as a trait, including artifacts or gerrymandered categories), whereas the second is too stringent (virtually ruling out behavior as a trait). The ACS strikes a balance between these extremes, providing a basis for researchers across many different areas of behavioral research to conceptualize

traits for study in both human-focused inquiry and non-human model organism research. Because ACS consists of a suite of analogies from the study of morphology that are neither necessary nor sufficient, it facilitates a shift away from the conceptualization of behavioral traits *in general* to the characterization of traits of interest according to specified criteria. This yields a basis for normative evaluations that rules some things in (e.g., learning and memory in fruit flies) and other things out (e.g., educational attainment in humans), with attendant consequences for different social and behavioral sciences.

2.2 Desiderata for an Evaluative Strategy

What counts as a behavioral trait for the purpose of scientific investigation? This broad question has been of interest to researchers for decades, but it has become increasingly pertinent in contemporary debates surrounding the use of Genome Wide Association Studies (GWAS) in the social sciences. Controversy surrounding Harden's 2021 book is one notable example. Marcus Feldman and Jessica Riskin criticized the book for promoting the use of GWAS to study "Social-Science-Relevant-outcomes" (e.g., traits like "open to experience", "grit", and "educational attainment"), which they claim are "interpretive descriptions, made of ideas, opinions, and practices, not molecules" (Riskin and Feldman 2022). To use GWAS to study educational attainment is "a category mistake" according to them, because educational attainment is not a trait for which there could be an underlying genetic architecture relevant to the differences observed in human populations. Underneath these criticisms is a conceptual issue: what distinguishes a bona fide organismal trait, thereby making it eligible for GWAS, from those that are not traits at all?

The reasoning behind this criticism from Feldman and Riskin is murky. One facet appears to be a concern about the criteria for traits being too liberal. If social science researchers consider educational attainment a trait, then what is to stop them from using GWAS to explore the genetic contribution to differences in made-up traits like "spurious reductionism" (the tendency to assume that all phenomena are reducible to nucleotides) (Riskin and Feldman 2022)? Another facet concerns how well researchers have conceptualized Social-Science-Relevant traits given their research aims and methods. For example, like any methodology, GWAS is a kind of hammer. As such, what counts as a nail? If the research aim is to understand the genetic basis of traits, then macular degeneration might be a clear nail, whereas schizophrenia might be nail-like, and educational attainment might not be a nail at all. One possible interpretation of the situation, suggested by some readers of Feldman and Riskin, is that behaviors are simply not traits, whereas morphological features clearly are. This interpretation seems overly strong; it worried readers concerned about deeming all behaviors as not traits. Other researchers charged Feldman and Riskin with painting with too broad a brush. To accept their account of the distinction would dismiss much of modern genetics, especially the ubiquitous studies of quantitative traits (Riskin and Feldman, 2022).

This debate is a microcosm of a larger issue that is present not just in behavioral genetics but also in social and psychological science research more generally. It serves as a useful starting point for examining how to approach answering the question of what a behavioral trait is. This entails investigating what researchers want from an account of behavioral traits, and crafting desiderata that are suited to actual scientific practice. In doing

so, one must also be aware of existing methods for conceptualizing behavioral traits and consider if they meet these desiderata. If not, one has an opportunity to suggest potentially better ways of evaluating behaviors as traits in the context of scientific research.

As illustrated in the opening exchange, there is a need for shared evaluative standards for what counts as a behavioral trait. But what do we want from such standards? One aim of scientific concepts, including categories of behavioral traits, is that they should be useful for empirical inquiry, which suggests that evaluative standards should be attuned to scientific practice. Second, as the concept of behavioral traits and its specific instantiations are used to do scientific work both across taxa and across disciplines, standards for evaluating behavioral traits should also be employable across diverse disciplinary communities and organisms of study. Third, any evaluative standards should be sensitive to the ongoing nature of scientific investigation. Thus, we would expect different standards of conceptualization for specific behavioral traits to be appropriate at different points in the investigative process. For example, we would not expect a newly observed behavioral phenomenon to be as well characterized as one that has been studied for a greater length of time and in more detail. Our standard for the latter would be higher than our standard for the former. The new behavioral phenomenon can be less clearly conceptualized as a trait early on in an inquiry and considered an object of study in need of further characterization. This is consistent with scientists deciding that a frequently studied, well-characterized behavioral phenomenon of interest may not be a trait.

Another desideratum researchers may want to identify traits is the ability to distinguish between measurement artifacts, proxies, and traits. For example, a

measurement procedure for a behavior might generate a behavior phenomenon purely as a consequence of undertaking the measurement (i.e., an artifact). Alternatively, we might measure one behavioral feature because it is correlated with another behavior that is difficult to assess reliably (i.e., a proxy). Researchers typically eliminate artifacts after they are identified but might exploit proxies regularly. This does not mean that discerning these distinctions is easy. Are ratios between the second and fourth digits of human hands artifacts, proxies for predicting personality or health, or a genuine trait (Leslie 2019)? Even if it is difficult to distinguish between artifacts, proxies, and traits, being able to do so is an appropriate standard for conceptualizing traits. For example, we should be able to say something about how proxies relate to traits. In other words, evaluative standards should speak to how behavioral traits are measured.

One might argue that a set of standards or evaluative strategy for behavioral traits must provide a basis for distinguishing *behavioral* traits from *morphological* traits. However, I resist adopting this as a desideratum. What counts as a behavior is not a settled matter (Zuk 2022; Zuk and Spencer 2020; Bergner 2016; Henriques and Michalski 2020; Levitis, Lidicker, and Freund 2009; van Duijn, Keijzer, and Franken 2006). In fact, some reject the distinction between behavioral and morphological traits altogether (Zuk 2022). At the very least, there are disciplines like molecular behavioral genetics or certain branches of evolutionary biology as well as organisms of study, such as microbes and plants, where this distinction is unclear or does little work. Therefore, a general evaluative strategy ought to focus on whether a phenomenon, especially a phenomenon considered to be a behavior by a discipline or for a specific organism being studied, is well

conceptualized as a trait without the additional burden of distinguishing between behavioral and morphological traits.

In summary, the four desiderata we have identified as comprising an adequate strategy for evaluating behavioral traits are:

1) Useful for doing scientific work - attuned to scientific practice
2) Applicable within and across disciplines and organisms of study
3) Responsive to changes associated with the nature of scientific inquiry
4) Distinguishes between behavioral traits, useful proxies, and measurement artifacts

With these desiderata formulated, it is helpful to introduce two behavioral phenomena where these evaluative standards might be applied.

2.3 Educational Attainment (in Humans), Learning and Memory (in *Drosophila*)

Viewing measurements of Social-Science Relevant outcomes as traits, such as educational attainment, social mobility, and socioeconomic status, is controversial. Yet many social science researchers use these outcomes as the basis of their investigations and employ methods for measuring them as traits. Educational attainment is often considered a cognitive behavior measured through subject surveys where individuals report either the number of years of schooling they have received or their highest educational certification (Gao et al. 2022; Yoshikawa and Asaba 2021; Martin et al. 2011; Davies et al. 2016; Rietveld et al. 2013; Lee et al. 2018; Davies et al. 2016; Hill et al. 2019; T.-T. Chen et al.

2024; Okbay et al. 2022). This way of conceptualizing educational attainment is used across a host of disciplines, including behavioral genetics, economics, public policy, social psychology, and epidemiology.

To compare educational attainment across different types of educational institution, both within and across national contexts, researchers utilize the International Standard Classification of Education (ISCE) ("International Standard Classification of Education - Statistics Explained," n.d.). To calculate educational attainment (EduYears), researchers first ascertain an individual's highest level of education as self-reported on a survey. Researchers then use the guide within the ISCE to convert the number of years or level of schooling across national contexts. For example, the United States and much of Europe differ in the typical length of their doctoral programs and the title of the degrees (3 years for a DPhil in the UK and 5 years in the United States for a PhD). Directly comparing two people with a doctoral degree from the different countries would yield a difference in the absolute number of years of education. However, using the ISCE, both individuals with doctorates would receive the same level: 8 EduYears.

Although the trait status of many behaviors in humans is contentious, this is not the case in model organisms. Importantly, conceptualization, individuation, and measurement of behaviors play critical roles in research across both contexts. Consider learning and memory in *Drosophila*. Researchers interested in uncovering genetic contributions to natural variation in learning and memory for populations of fruit flies must be able to define their character of interest and its variation. Through a variety of empirical investigations, including experimental intervention strategies, researchers have gained information about

learning and memory phenomena in *Drosophila*. This work has included making distinctions about these phenomena based on the temporal persistence of the behavior, resulting in distinctions like short-term memory, anesthesia-resistant memory, and long-term memory. This knowledge informs studies on particular kinds of learning and memory, such as associative olfaction learning. Variation in associative olfaction learning and memory is assessed using behavioral assays that are developed to be sensitive to the difference between different kinds of memory (Wang et al. 2023). Researchers have leveraged these distinctions about learning and memory to investigate the relationship between natural genetic polymorphisms at the *foraging* (*FOR*) locus and associative olfaction learning (Mery et al. 2007).

To measure associative olfaction learning, researchers must engage in a two-part process. First, they need to build an association to a scent. Researchers use an aversive olfactory conditioning assay to compare learning performance between flies homozygous for *forR* (roving) and *forS* (sitting) (Mery et al. 2007). The protocol for the learning assay exposed fifty adult flies (3-5 days old), homozygous for either *forR* or *forS*, to a classical conditioning method where the flies associated a scent (either octanol or methylcyclohexanol) with a mechanical shock through standardized sequences of conditioning. The flies were exposed to one scent for 30 seconds, while simultaneously being shocked for one second at five-second intervals. Flies were given a 60-second rest period followed by an exposure to the other scent for 30 seconds without being shocked. This was followed by a 60-second rest period. Some fly cohorts received only one round

of conditioning; others received multiple rounds without an intervening rest period; a third group received multiple rounds of conditioning separated by 20-minute rest intervals.

The second step in measuring associative olfaction learning is to assess the persistence of this association. Differences in behavior after conditioning were evaluated in a T-maze using three different character states for memory: short-term, long-term, and anesthesia-resistant. Within the maze, flies were exposed to two air streams, one carrying the first scent and the other carrying the second scent. Two samples of flies, identical except for the scent they were conditioned to avoid, were then allowed to "choose" between (i.e., fly toward one of) the two odors for 60 seconds. Flies that did not move from the entrance of the T-maze were excluded from the study. To derive a single value of the memory index, researchers calculated the difference in the proportion of flies choosing octanol in these two samples.

In the cases of both educational attainment in humans, and learning and memory in fruit flies, researchers are investigating behaviors and consider them traits. However, within the broader research community, the claim that educational attainment is a behavioral trait is highly controversial. On the other hand, the claim that associative olfaction learning is a behavioral trait in *Drosophila* is widely accepted. Should both be considered behavioral traits? If so, why? To answer these questions, we need an account of what counts as a behavioral trait.

2.4 Conventionalism and Mechanism as Evaluative Strategies

There are two commonly deployed evaluative strategies for giving an account of what a behavioral trait is: conventionalism and mechanism. These two strategies are idealized depictions of actual (though often implicit) methodological approaches to the identification of behavioral traits. Any given researcher may have a hodgepodge of commitments regarding how to define a behavioral trait, or they may align more clearly with one of these two strategies. I reconstruct the evaluative strategies for trait conceptualization by attending to the aspects on which researchers focus when measuring a trait. How researchers measure a behavior of interest signals (in part) how they conceptualize it. After making the standards of evaluation for each strategy explicit, I argue that the standards of evaluation for these existing strategies either require too little (anything which can be measured qualifies as a trait) or too much (one must know much or all the underlying causal mechanism).

2.4.1 Conventionalism

Conventionalism, as an evaluative strategy, proposes that what counts as a behavioral trait is simply that it can be measured. This strategy has roots in the operationalization of measurement, where to measure a phenomenon of interest, such as temperature change or height, one must first go through a process of developing a procedure for capturing the phenomenon in standardized terms (e.g., Fahrenheit/Celsius, meters) (Bridgman 1927;

Boring 1945; Skinner 1984; Reichenbach 2012). Thus, being able to measure behavior in a standardized fashion is sufficient for it to be defined as a behavioral trait. Philosopher Sean Valles has argued for a version of conventionalism in his work on trait validity and trait utility (Valles 2013). He appeals to Dobzhansky's 'dismissive' view of traits: "The isolation of 'traits' is, then, only a semantic device valid and useful in recording our observations" (Dobzhansky 1956, 342). This view foregrounds the active role of human researchers in carving up organisms into identifiable traits. Valles argues that the ability to measure an aspect of an organism in a standardized way is sufficient for it being a valid trait (Valles 2013).

Educational attainment in humans and olfactory learning and memory in fruit flies can both be measured. Therefore, according to conventionalism, both are valid behavioral traits. However, this strategy does nothing to alleviate the concern of Feldman and Riskin that some measured traits could result in arbitrary designators that are "a reflection of its creator's radically subjective view of things" (Riskin and Feldman 2021, 5). Are these two traits merely defined arbitrarily?

Although the process of measuring learning and memory in *Drosophila* involves carving up the behavior of a whole fluid organism into diverse aspects, I argue that this process is far from arbitrary or radically subjective. In *Drosophila*, trait measurement occurs by subjecting the organism to different environmental conditions (classical aversion conditioning) and then observing the organism's response. It is worth noting that in *Drosophila*, the behavior of interest is temporally bounded. After being conditioned, fruit

flies are assessed at strict time intervals. The assessment itself takes place within a tight window—just 60 seconds. Furthermore, the apparatus (the T-maze) is standardized. One could go to any lab, test any group of flies, and be reasonably certain that the measurements would say something about the trait of interest, in the context of the surrounding environment.

Similarly, measures of educational attainment are not arbitrary or radically subjective. By utilizing the ISCE to calculate EduYears, researchers use a consistent method for assessing a given participant's level of education across national contexts. The criteria for categorizing each of the eight levels attempt to group kinds of national systems of education based on their structure and content. However, there are important aspects of educational attainment not captured by the EduYears measurement methodology. These include cultural factors, such as financial means, political structure (whether there is government mandated military service), SAT/ACT preparation, degree of support from the family, and the multi-faceted quality of education. All these aspects can affect an individual's experience and persistence within a given educational system. Yet failure to capture these aspects does not mean the measurements are arbitrary. It could be that researchers are focusing on a subset of specific aspects of educational attainment given their research aims.

However, there is a more pressing issue for conceptualizing educational attainment as a behavioral trait. Whereas many researchers refer to educational attainment as a trait (Martin et al. 2011; Lee et al. 2018), others view it as a proxy for some underlying feature, such as cognitive ability (Berry, Gruys, and Sackett 2006; Rietveld et al. 2014). Even

within the same study, educational attainment can be labeled as both a trait and a proxy (Lee et al. 2018). Proxies serve important roles in scientific investigation. They allow researchers to gain traction on otherwise inaccessible (or difficult to assess) phenomena. For example, one may be interested in variation in cognitive ability but measure it with the proxy of performance on IQ tests. However, administering and analyzing direct measures of cognitive ability is time and labor intensive. This information is not collected by most DNA biobanks and is therefore unavailable within the data sets used by those studying relationships between genomic variation and educational attainment. Utilizing only those datasets having direct measures would severely limit the sample size of studies and greatly increase their costs and time to completion. Instead, researchers may choose to use a proxy, with the goal of getting at cognitive ability indirectly. Researchers operating multimillion participant studies use educational attainment as their proxy. Intuitively this makes sense. Surely cognitive ability and educational achievement must travel together to some degree (i.e., are in some sense correlated). However, educational attainment also has been used as a proxy for financial and social status (Liu 2019; Morris et al. 2020). So, are researchers tracking cognitive ability, or socioeconomic status, or both?

Regardless, a proxy for a trait ought to have a different status than the trait itself because it serves a different role in the investigation. Yet conventionalist strategies have no principled way to distinguish between proxies for traits and traits themselves. For conventionalists like Valles, trait validity and trait utility are separate issues. Questioning whether educational attainment is a useful trait in scientific research has no effect on its status as a valid trait. Yet this is exactly the question at issue in current social science

debates. Additionally, because proxies are inherently useful, there needs to be some explicit account of how the use of a proxy links up to the trait of interest (e.g., cognitive ability). Therefore, conventionalism fails to meet a key desideratum: the ability to distinguish between a behavioral trait, proxy, and artifact of measurement. Furthermore, it is unclear whether conventionalism can answer the question of trait validity independently of trait utility.

2.4.2 Mechanism

A second potential strategy for defining behavioral traits is through knowledge of their underlying molecular and cellular mechanisms. A strong version of this strategy could draw inspiration from recent theoretical work about how to understand morphological traits: Character Identity Mechanisms (ChIMs) (DiFrisco, Love, and Wagner 2020; DiFrisco, Wagner, and Love 2023). The ChIM framework is specifically formulated to assist with identifying homologous structural characters, traits that are shared between two taxa due to common descent. Under this account, one can identify a homologous trait even when it can be multiply realized and exhibit diverse phenotypes. To do so one must discover the identity mechanism shared by homologues. For example, consider the tetrapod limb. Traditional evolutionary biology identifies it as a homologue by the mere fact of common descent. However, the ChIM framework offers a specific explanation: the tetrapod limb has an underlying identity mechanism whether it is instantiated in a dolphin's flipper, a horse's forelimb, or a human hand. The ChIM framework is also useful for ongoing investigation because aspects of the identity mechanism can be elucidated and

then used to trace homologies in comparative biology for diverse taxa. Mechanistic strategies, more generally, offer a clear basis for what counts as a character that fulfills several of our desiderata: useful for doing scientific work and attuned to scientific practice; applicable within and across disciplines and organisms of study; and responsive to changes associated with the nature of scientific inquiry.

However, utilizing a mechanistic strategy for trait conceptualization necessarily relies on having access to key aspects of the mechanism that gives rise to a trait. Researchers working in model organisms often have secured experimental access to at least some of these underlying mechanisms, although this typically requires tremendous time and effort. Given the structure of these communities, researchers can draw on shared databases, breeding lines, genomic and transcriptomic data, as well as standardized intervention methodologies (Ankeny and Leonelli 2021; 2011). Furthermore, interventionist methods are less ethically and practically constrained in model organisms. Access to these kinds of resources allows researchers interested in a given trait to draw on this wealth of information when elucidating mechanistic aspects underlying it. This is well illustrated by our example of associative olfaction learning and memory in fruit flies; researchers drew on knowledge of the underlying mechanism to distinguish between short-term memory, long-term memory, and anesthesia-resistant memory. Decades of experimental and observational studies have yielded mechanistic knowledge of learning and memory phenomena in *Drosophila* across multiple levels of organization. Researchers have identified anatomical structures centrally involved and are even able to distinguish between anatomical structures that are used in different kinds of learning and memory (e.g.,

appetitive vs. olfactory) (Davis 2023; Busto, Cervantes-Sandoval, and Davis 2010). At a finer grain, researchers have learned about the various kinds of cells present within these structures, including their location and orientation with respect to other structures and cell types (Schwaerzel et al. 2003).

A mechanistic strategy for defining behavioral traits has at least two benefits. The first, illustrated by the ChIM framework, is the ability to track a shared feature across taxa by identifying relevant aspects of shared molecular and cellular mechanisms. The second, illustrated by learning and memory in *Drosophila*, is the ability to distinguish between observationally similar traits by appealing to differences in the underlying mechanism of the traits. These benefits accrue only if researchers have access to aspects of an underlying mechanism. Having this access for a given behavioral trait is often deeply informative, but the general limitations on access to these details greatly limits the applicability of the strategy. Some domains of research have greater access to information about mechanisms than others. Researchers interested in the behavior of model organisms have access to vast numbers of studies, shared resources, and experimental techniques. Our best characterized behavioral mechanisms come from research performed in model organisms. Meanwhile, vast areas of behavioral research, particularly human-oriented work in social, behavioral, and psychological sciences, are severely handicapped.

In response to this handicap, investigators operating in human research contexts often treat behaviors heuristically as traits, in order to gain insight into particular mechanisms. This is especially prevalent in areas of social sciences where researchers want to understand the relationship between expression of a behavior and elements of the

environment for the purpose of intervention. Here, researchers may temporarily ignore the status of a behavior's "trait-ness" to facilitate insight into mechanistic detail. However, demanding that researchers understand underlying mechanisms to provide a working account of behavioral traits disregards scientific practice. Scientists often choose to partition or selectively ignore certain areas of investigation to make progress on a given issue. For example, Charles Darwin did not understand the mechanisms of inheritance assumed in his theory of descent with modification. However, we collectively describe Darwin's theory as an advance in scientific understanding. This process has been called "black boxing"—scientists deliberately set aside aspects of phenomena which they do not understand to make progress on other aspects (Nathan 2021). Researchers studying behavior in humans often engage in black boxing out of necessity because their investigative methods are constrained by practical and ethical concerns.

Although a mechanistic strategy may add a sophisticated level of clarity and detail to an account of what behavioral traits are, it is burdensome. An extraordinary amount of detail must be known about the mechanism of a potential trait for researchers to ascertain if an observed behavior meets the criteria. Few domains of behavioral research have access to such information, especially those focused on humans. This greatly limits the areas of study that could confidently identify behavioral traits, violating the desideratum that an evaluative strategy should be applicable across diverse taxa and domains of research. Furthermore, relying on a mechanistic strategy for behavioral trait conceptualization would greatly constrain researchers' ability to investigate important domains, including human

behavior. Thus, the mechanistic strategy violates the desideratum that trait conceptualization be useful when humans are a primary object of study.

2.5 An Analogical Contextual Strategy

2.5.1 Background

Because previous philosophical work on evaluating behavioral traits was motivated by concerns about homology and evolutionary novelty (e.g., identifying homologues across a phylogeny), the little work that has been done is largely embedded in this literature and responding to it (Brown 2014; Valles 2013; Longino 2013). For example, the philosopher Rachael Brown draws on work for evaluating homologous morphological traits using traditional criteria of homology (including relative position, special quality, and connection by intermediates) and extends them to the evaluation of behavioral traits.[6] Special quality criteria and connection by intermediates operate as necessary and sufficient bases for determining homologous behaviors within a lineage for Brown. For example, in order for a behavior to be homologous under the special quality criterion, one looks for a distinctive element of the behavior that is thought to be non-adaptive and similar to others in the organism's lineage (Brown 2014). If a distinctive non-adaptive element cannot be found, then the behavioral trait is considered non-homologous.

Tinkering with traditional criteria to engender applicability to behavior does offer promise as at least part of an evaluative strategy, but on its own it is limited in scope. This

[6] This strategy has also been pursued by biologists interested in behavioral traits (Wenzel 1992).

evaluative strategy is best suited for identifying homologous and evolutionarily novel behavioral traits, mostly of interest in evolutionary biology. Evaluative strategies that prioritize these research questions offer less assistance for behavioral traits generally, and especially in the non-comparative context of human biology. In fact, this fixation on function, homology and novelty motivated Valles to develop a conventionalist account that could be applied in domains like public health. However, as we have seen, conventionalism has its own pitfalls. Therefore, we need a different and more general strategy for evaluating behavioral traits that can be used across disciplines and research programs, as well as being sensitive to the difference between traits, proxies, and artifacts of measurement.

2.5.2 Articulating a new strategy

To meet this need, I offer a new general strategy for evaluating behavioral traits that consists of a cluster of criteria, each of which is neither necessary nor sufficient. Together, however, they meet the desiderata described in Section 2.2. The proposed criteria—temporal boundaries, relationship to function, regularity of occurrence, and appearance at specified life history moments—are inspired by analogies to the kinds of considerations researchers raise when characterizing morphological aspects of an organism. However, they have been adapted to better fit the needs of behavioral researchers. For example, unlike the strategy and its criteria suggested by Brown, evaluations using these criteria are not necessarily discrete categorical assessments (i.e., this is a trait, that is not a trait). Rather, these criteria work together to yield qualitative evaluations that are especially sensitive to the usefulness of a particular trait with respect to certain research aims in a specific context.

They yield both singular and comparative evaluations (e.g., this trait is sufficiently characterized for this research purpose, or that trait is better characterized for a specific purpose than another trait). In Section 2.6, I explore how this strategy can be used to evaluate behavioral traits and what these evaluations tell us about future research. But first I provide a description of the proposed criteria.

2.5.2.1 Temporal Boundedness

Just as morphological features manifest within space, behavioral features do over time. Part of the basis for accepting the forelimb as a morphological trait is that we can spatially circumscribe it; this includes its relation and position with respect to other morphological features. Our specification of the spatial boundaries of a given character lends credibility to it being a morphological trait. A researcher's designation of a behavioral trait within a defined temporal dimension can play a similar role. Researchers acknowledge that behaviors exhibit a great degree of variation both within and among taxa, especially along the temporal dimension. For some behaviors, it may be easier to identify clear start and end elements, as in the case of mating behaviors (e.g., courtship songs). Defining the temporal boundaries of behaviors is often not enough to address certain research questions. Researchers may also need to characterize behaviors according to a sequence of actions through time. Similar to dissecting a morphological character into its component elements, researchers decompose a behavior into smaller segments or routines and track their occurrence in a particular sequence.

Importantly, different kinds of behaviors exhibit greater and lesser degrees of ritualization, or variation between individuals within a particular taxon, along temporal

dimensions. Therefore, the way that researchers operationally define behaviors in contexts of measurement plays an important role in characterizing a given behavioral trait. In this sense, researchers' own awareness of the difficulty of measuring behavioral traits with respect to a temporal boundary may serve as an important guidepost when making judgements about the "trait-ness" of a behavior. In extreme cases, inattention to the temporal dimension of a behavior may itself signal an under-characterized trait. Notice that to evaluate the temporal boundedness of a potential trait one must look to scientific practice, specifically measurement practices. In doing so, this criterion fulfills a key desideratum of a strategy for characterizing behavioral traits—being attuned to scientific practice.

2.5.2.2 Regularity of occurrence

Just as morphological features that appear with greater frequency are better characterized than those that appear less frequently, the same should be expected for behavioral features. However, this does not mean all behavioral features will follow this pattern. Instead, this criterion operates as a rule of thumb (i.e., neither necessary nor sufficient, but heuristic). Consider snout morphology in the Talpidae. In this family of small moles, most members have a standard snout. However, within this family, one tribe has an extant species (the star-nosed mole) with an unusual snout containing a specialized sensory structure. Scientists are working to characterize this morphological feature, but early on it was far less characterized than the standard snouts of other moles in its family.

We can expect many behaviors to follow a similar pattern, in part because behaviors that occur more frequently provide more opportunities for study. In contrast, behaviors that

occur less frequently are more difficult to characterize. In fact, the behaviors commonly referred to as "well-characterized" tend to be behaviors that occur frequently, such as mating, foraging, parenting, communication, and combat. These kinds of behaviors manifest both within a given taxon and across a diverse set of taxa. Furthermore, within a given taxon these traits often exhibit a predictability of occurrence in response to specific stimuli. For example, *Drosophila* mating behavior is well characterized in part due to its frequency of occurrence; almost all relevant individuals exhibit this behavior multiple times at a particular stage of their life history in response to specific stimuli (see below) and therefore it occurs with high predictability. Although one can expect violations of this criterion, it provides a key point of traction for characterizing many behaviors. This criterion promotes and even requires an evaluator to look both within and across taxa to assess the regularity of occurrence of a potential behavioral trait, fulfilling the second desideratum for an evaluative strategy, applying across diverse organisms of study.

2.5.2.3 Appearance at specific life history stages

Just as some morphological features are tied to particular life history stages (e.g., secondary sexual characteristics), so are many behaviors. Researchers know that most mating displays are situated during adult sexual maturity, while filial imprinting occurs during the perinatal sensitive period (e.g., ducklings following their mother). This criterion, which is related to temporal boundedness, draws special attention to developmental aspects of behaviors. Behaviors of organisms change over time; both the degree and manner of change may differ. Having an awareness of change in behavior over an organism's development can facilitate fine grained observation and study, advancing the characterization of a given

behavior. Consider again imprinting behavior in birds. The life stage of the organism is a critical aspect of the characterization of this behavior. All other things being equal, imprinting will not occur if the bird is at a different developmental stage, and imprinting behavior eventually disappears as the bird develops. The variation and persistence of a behavior across life history stages is also a dimension to consider when evaluating whether a behavior is well characterized for a specific context of inquiry.

2.5.2.4 Relationship to Function

Many have noted that discussions of traits and functions often travel together (Valles 2013; Gould and Lewontin 1979; Sterelny 1988; Violle et al. 2007; Ariew, Cummins, and Perlman 2002). However, the extent to which these concepts should travel together is debatable. For example, teasing them apart is crucial for investigating the homology of morphological features. Yet some see these concepts as deeply linked: "All traits are potentially linked to fitness and thus are functional by definition" (Sobral 2021, 675). For others, they can be separated (Valles 2013). Function and adaptation might only become relevant when researchers are interested in a trait's origin or evolutionary utility. While I agree with Valles that having a function or being adaptive should not be a requirement for being a behavioral trait, questions of function and adaptation do play an important role in the process of characterizing behavioral traits. Importantly, this extends to multiple ways one can talk about function, including activity, causal role, biological advantage, and selected effects (Wouters 2004).

Attending to questions of function helps researchers characterize behaviors. For example, concentrating on activity draws attention to movement, vocalizations, and color

or textural changes that occur during or signal the onset of a behavior. This may include reference to specific structures and how their relationships to one another change during a behavior. When rattlesnakes are stressed, they assume a particular "defensive posture" where the body is coiled and their rattle is raised, vibrating back and forth. Concentrating on a behavior's biological or causal role draws attention to its connection with self-maintenance, growth, development, and reproduction, as well as raising questions about underlying mechanisms (e.g., muscle twitching that sustains rattling). For the snake, rattling behavior is part of its defense repertoire. Attending to a behavior's biological advantage prompts a researcher to consider a kind of counterfactual in which the trait in question is compared to alternatives to assess the utility of the actual trait as beneficial, neutral, or detrimental. Researchers may be interested in the utility of different behaviors snakes exhibit when threatened, including fleeing and striking. Attending to the effects for which a behavior seems likely to have been selected in the past encourages researchers to connect a behavior with its evolutionary history. Researchers may investigate the evolution of striking or rattling behaviors in rattlesnakes by examining the presence or absence of the behavior in extant relatives across a phylogeny or by the kinds of variables that influence the behavior.

Significantly, the aspects and conceptions of function determined to be most relevant depend on the aims of researchers. Those interested in the genetic architecture of a behavior for the purpose of intervention are likely to characterize a behavioral trait in terms of its causal role (e.g. what the behavior causes) rather than selected effects. And failing to characterize a behavior in terms of one or more aspects of function does not mean

the behavior is not a well characterized trait. Rather, this criterion evaluates whether researchers have engaged with notions of function relevant to their aims of inquiry.

The relationship to biological function and its ties to the other three proposed criteria further cement the Analogical Contextual Strategy's ability to meet our desiderata. The relationship to biological function plays an important role in ensuring the strategy applies both within and across diverse taxa and organisms of study. Many different disciplines (including genetics, molecular biology, anatomy, and biochemistry) are involved in uncovering behavioral mechanisms. Additionally, many taxa, especially model organisms, are studied to elucidate a general behavioral mechanism for frequent behaviors. Therefore, the relationship to biological function clearly ensures that a strategy meets the desiderata of applying within and across organisms and disciplines of study.

Relationship to biological function also plays a critical role in distinguishing between behavioral traits, useful proxies, and artifacts of measurement. A common aspect of ascribing (or not ascribing) biological function is understanding the relationship between one part of an organism and another. This happens to be a key feature in distinguishing between a trait, proxy, and artifact. Therefore, attending to a behavior's relationship to biological function also fulfills this desideratum. Finally, relationship to biological function foregrounds the aims of the researchers in the assessment of potential behavioral traits. In doing so, the relationship to biological function allows the framework to be sensitive to changes associated with the nature of scientific inquiry.

What we now have in hand is a general strategy (the Analogical Contextual Strategy) for evaluating behavioral traits. It consists of a cluster of criteria that are neither

necessary nor sufficient—temporal boundaries, relationship to function, regularity of occurrence, and appearance at specified life history moments—yet these criteria work together to yield qualitative evaluations that meet the articulated desiderata and are especially sensitive to the usefulness of a particular trait to achieving certain research aims. The proposed criteria are inspired by analogies with the kinds of considerations at play in the characterization of morphological traits and therefore have prima facie warrant. Now we turn to exploring how this strategy can be used to evaluate behavioral traits in scientific practice.

2.6 Scientific Payoff

It is time to return to our starting questions: Is associative olfactory learning and memory in *Drosophila* a behavioral trait? And can educational attainment in humans also be characterized as a behavioral trait? Through an evaluation of these two behaviors, I aim to demonstrate how the ACS can be applied across diverse disciplines and organisms of study to yield evaluations relevant to ongoing research. Let us begin with associative olfactory learning in *Drosophila*.

As described in Section 3.2, the associative olfactory learning in *Drosophila* is temporally bounded. Researchers have methods for measuring it that include distinct time intervals in both the conditioning phase of learning and in the assessment criteria. This method of carving up the behavior temporally is standard in the field, with multiple laboratories and research groups using the same method.

In terms of regularity of occurrence, associative olfactory learning and memory behaviors occur frequently. Most individuals in the population exhibit the behavior, and it is far from unique to *Drosophila melanogaster*. Similar behaviors are observed in other insects, mammals, and birds (Caldicott et al. 2023; H.-L. Chen, Chen, and Huang 2022; Nelson, Slater, and Hauber 2017; Faber, Joerges, and Menzel 1999). Furthermore, researchers have a strong understanding of how environmental variables can influence variation in the exhibited behavior and its acquisition.

Learning and memory behaviors are routinely characterized in terms of life history stage. In addition to research focused on associative olfactory learning in adult flies, studies have also been performed in larvae. Fruit flies experience significant growth and development between larval and adult life stages. The assessment and conditioning methods differ between these stages. For example, researchers use gustatory reinforcement in larvae instead of adverse shock conditioning in adults (Scherer, Stocker, and Gerber 2003). These differences in both behavioral conditioning and assessment may map on to the major changes facilitated by growth and development in these life history stages. The differences in methods of assessing the behavior may be attuned even more narrowly to finer grained life history stages. More generally, detection of and navigation towards attractive odorants are critical at both life stages because fruit flies consume nutrients in both their larval and adult life stages (something not true of all insect species).

Finally, associative olfactory learning and memory are together characterized in terms of relevant notions of function. In general, associative olfactory learning and memory is connected to important aspects of biological function, including foraging behavior and

harm reduction. However, this alone is not sufficient to judge whether the behavior is well characterized, because the aims of specific researchers must be considered. In the study described in Section 3.3, researchers were investigating associative olfactory learning and memory to identify genes that contribute to natural variation in cognitive behavior and to further elucidate how differences may evolve in populations and species (Mery et al. 2007). While these are distinct but related aims, I focus here only on the identification of genes that underlie natural variation in the trait. For a behavior to be well characterized according to this aim, researchers must understand the behavior's mechanism. Researchers do have the behavior sufficiently characterized in terms of its mechanism to be able to quantify variations in the behavior. Thus, across most if not all the criteria, associative olfactory learning and memory in *Drosophila* is well characterized. There is ample justification for learning and memory in *Drosophila* to count as a behavioral trait generally, and specifically in domains concerned with the genetic underpinnings of differences in the trait.

How well does educational attainment in humans meet the criteria for being considered a trait as evaluated by the Analogical Contextual Strategy? Educational attainment is not well characterized in terms of temporal boundaries. While other kinds of cognitive behaviors like speech and language acquisition are geared towards assessing students' skills given their developmental stage, educational attainment is not. Rather educational attainment is assessed across a person's lifetime as a retrospective summary or at the time the survey is answered. Most often, educational attainment is assessed when subjects are in their thirties, which is considered past the age that higher education occurs. Yet, one can acquire higher education at almost any age, although it becomes increasingly

less likely as individuals get older. Furthermore, the common levels of education available to participants vary according to the decade in which education is assessed. For instance, many women and people of color were either prohibited from higher education or simply lacked the opportunity to enroll in higher education until recently. Additionally, the education systems reflected in educational attainment measurements are institutions that arose within the last 200 years and were standardized within the last 100 years. Thus, differences in educational attainment are not comparable across time periods.

The characterization of educational attainment in terms of regularity of occurrence is questionable. It measures the education of an individual across their lifetime. However, educational attainment is not instantiated multiple times across one's life. Therefore, within humans, it can only occur once for an individual. And quite evidently, educational attainment cannot be observed across taxa. The system of education as measured in years and progress in attainment of degrees is constructed for humans. Therefore, humans may be the only taxon that can exhibit educational attainment. Other organisms, such as sport horses and guide dogs, are subject to formalized systems of training, but it is uncontroversial to say these are not comparable educational systems (for many reasons) and therefore do not warrant a kind of cross- taxa comparison. Importantly, being observed only within one taxon does not rule out educational attainment as a trait. However, when considering issues of temporal boundedness, educational attainment is found only in humans and only appeared in the last 200 years. The rationale for seeing educational attainment as a trait is greatly reduced.

Educational attainment as a trait is not connected to a specific life history stage. It is only assessed in adulthood. The individual levels of EduYears often have a loose correspondence to chronological age, with lower levels associated with younger developmental stages and only higher levels associated with adulthood. However, this is a relatively loose connection.

Educational attainment is not well characterized in terms of biological function and causal role. Researchers have made general claims that educational attainment is associated with mate choice, health outcomes, economic stability, and reproduction, social standing, and class (Gao et al. 2022; Boggess 1998; Eaves and Hatemi 2011; Miech and Hauser 2001; Filmer and Pritchett 1999). However, how these associations are formed (including whether they might be spurious) is still hotly debated. Including educational attainment in particular research agendas or studies can be productive so long as researchers recognize limitations in the way the behavior has been conceptualized. For example, using educational attainment in social science research to track the impact of economic or social inequality may not pose an issue. Social science research is also not committed to a deep notion of biological mechanisms; researchers can be somewhat agnostic about mechanisms underlying the development of behaviors. However, considering educational attainment to be a trait for the purpose of exploring how genetic variation is related to behavioral variation is a mistake. Behavioral genetic research has an implicit commitment to mechanism—there must be some pathway that connects variation in genes to variation in the behavior. In the absence of conceptualization regarding molecular or cellular mechanisms for educational attainment, making the connection from genes to behavior is

premature. As it stands, educational attainment is poorly characterized across the relevant criteria for consideration as a trait in many domains of study, especially those in which a commitment to mechanism is necessary.

However, recognizing a behavior as poorly characterized with respect to specific parameters does not mean research on these features cannot continue. Rather, it is an invitation to engage in research that facilitates conceptualization along these lines. One consequence of employing the ACS is that it reveals specific areas where conceptualization of a potential trait can be improved, suggesting that researchers should prioritize this kind of research. Educational attainment is considered a general cognitive behavior interrelated with better conceptualized behaviors like early childhood first language acquisition and spatial reasoning. It may be fruitful to investigate other behaviors that are highly relevant to educational attainment and more deeply rooted in the above criteria to better characterize educational attainment. As shown above, although educational attainment is a phenomenon that we can investigate, we need to hold on to it lightly. Educational attainment serves an important investigative role as a proxy, and future research may require that we reevaluate the status of educational attainment as a proxy or trait. Thus, recognizing when a behavior is poorly characterized as a trait can be an invitation to change investigative approaches to better track the above criteria, which has the consequence of augmenting our understanding of less well-characterized features.

The ACS criteria serve as a framework for areas of research to better investigate behavioral traits. The strategy acknowledges the importance of mechanisms, while also facilitating research in more restricted domains such as human behavior. Furthermore, this

strategy is more permissive when it comes to human behaviors or species-unique behaviors. For example, early childhood first language acquisition meets the standard of the criteria for this strategy, even though the mechanism, especially its molecular features, are still largely unknown. However, this strategy promotes cross-taxa investigations because it highlights mechanisms underlying behavior as important features and points to the frequency of the behavior across taxa as an important consideration. This, in turn, may motivate those in more methodologically restricted fields to examine research performed in non-human model organisms. Finally, this strategy highlights the relationship between trait conceptualization and research methodology, especially regarding issues of measurement.

The ACS presents a preliminary basis for researchers across many different areas of behavioral research to conceptualize traits for study in both human-focused inquiry and in model organism research. Importantly, this strategy may be of particular relevance to disciplines like behavioral genetics that have been criticized for conceptualizing traits simply on the basis of their being measurable (i.e., conventionalism). Some of these critiques challenge the idea that behavioral traits can, in general, be clearly conceptualized. By adopting this strategy, one can shift the conversation from the conceptualization of behavioral traits in general to the characterization of specific traits of interest according to specified criteria. This yields a basis for normative evaluations, suggesting that learning and memory in fruit flies counts as a well-characterized behavioral trait, whereas educational attainment in humans does not. This distinction brings with it attendant

consequences for different social and behavioral sciences, especially in terms of what lines of investigation should be pursued in order to make empirical progress.

2.7 Conclusion

Within this chapter I sought to answer the question of what counts as a behavioral trait for the purpose of scientific investigation. In Section 3.2, I leveraged contemporary debates in behavioral genetics to identify key desiderata for evaluating behavioral traits. These were: useful for doing scientific work and attuned to scientific practice; applicable within and across disciplines and organisms of study; responsive to changes associated with the nature of scientific inquiry; and, able to distinguish between behavioral traits, useful proxies, and measurement artifacts. In Section 3.3, two cognitive behaviors, educational attainment and associative olfactory learning and memory, were introduced as examples of traits that are used in research to study the connection between genetic variation and behavioral variation in human and non-human organisms. In Section 3.4, I evaluated these behaviors using two approaches to behavioral traits: conventionalism and mechanism. In doing so, I demonstrated that both approaches failed to meet the desiderata. In Section 3.5, I developed a new strategy—the Analogical Contextual Strategy—for evaluating behavioral traits by drawing analogies to how researchers characterize morphological traits. This approach offered a cluster of criteria that are neither necessary nor sufficient but serve as a heuristic guide for identifying behavioral traits for the purpose of research: temporal boundaries, relationship to function, regularity of occurrence, and appearance at specified life history moments. In Section 3.6, I evaluated educational attainment and associative olfactory

learning and memory using this new strategy. By adopting the ACS, one can shift the conversation from the conceptualization of behavioral traits in general to specific traits of interest according to specified criteria. ACS can be used to justify normative evaluations, providing a strategy for distinguishing between well-characterized traits and those that are not. These normative evaluations are of consequence for social and behavioral sciences because they encourage improved investigative approaches to behavioral traits.

Chapter 3: Methodological Shortfall

3.1 Introduction

There is a common image of science, a picture of what makes scientific knowledge distinctive. This image includes expectations about how science is meant to operate. For example, one common expectation is that scientific theories are changed or abandoned when faced with negative evidence. Another expectation is that research questions are structured to allow researchers to consider different possible answers. One role for philosophy of science has been to interrogate patterns of operation in the sciences that seemingly deviate from these common expectations. That is, philosophers are very interested when science operates differently from what many people would expect.

Consider several examples. Thomas Kuhn investigated why researchers hold on to theories even in the face of accumulating empirical anomalies (Kuhn 1964). Imre Lakatos explored the conditions for scientists to continue pursuing a research program, even when it wasn't receiving experimental confirmation (Lakatos 1970). Elisabeth Lloyd analyzed why some scientists seem unable to countenance different possible answers in the face of mounting negative evidence (Lloyd 2015). Studying episodes of science operating differently from what we expect offers unique windows for philosophers to say something about how science works. In some cases, the conclusion is that we *ought* to expect this pattern of operation. This would prompt us to update our understanding of science and see these formerly puzzling episodes as evidence of how it should operate. In other cases, counterintuitive activities could signal that science is going wrong. Recognizing these

instances can assist philosophers and scientists to identify corresponding solutions to help get a science back on track. Although previous philosophical work has focused intensely on the ways theoretical commitments and research questions can lead science to operate in counterintuitive ways, little attention has been paid to the role scientific methods may play in similar episodes. What about instances where scientists continue to use a set of methods, even though these methods yield diminishing returns or have not led to empirical progress?

In this chapter, I examine one type of counterintuitive activity in science related to the role of methods— when research communities continue to pursue research approaches despite their inability to test central hypotheses. I refer to this pattern of activity as "methodological shortfall." I argue that methodological shortfall occurs when research communities concentrate their research programs around central questions, but their available methods are constrained in their ability to answer these questions unambiguously. In other words, a research community's reach exceeds its grasp. The available methods seem to provide some answers to questions of interest, but these answers may be misleading, inaccurate, or wrong. If faced with methodological shortfall, researchers have two response options: (1) work with available methods to try to overcome the existing shortfall (including elaborating on existing methods), or (2) adopt different approaches not subject to the shortfall. This second option often involves learning new methods and redirecting resources. There may be a high upfront cost of effort, money, and time with the second option, but it may be a necessary, and eventually more empirically fruitful, strategy.

My argument will proceed as follows. First, I will distinguish methodological shortfall from other existing explanations of science acting counterintuitively. In the case of methodological shortfall, not only is science operating counterintuitively, but the science is going wrong. I then use contemporary behavioral genetics (BG) as a case study to demonstrate how the concept of methodological shortfall can help us make sense of ongoing complaints about BG research undertaken on humans. Analyzing these issues through the lens of methodological shortfall reveals two paths forward for human BG research: (1) continue to use observational studies like Genome Wide Association Studies (GWAS) and Twin Studies, pursuing incremental improvements such as increased sample sizes, or (2) adopt experimental methodologies that can only be used in model organisms, which means approaching their object of study (i.e., behavioral differences) in an indirect way. After characterizing these two paths, I journey down each one to highlight how that choice impacts epistemic progress. Some researchers in BG have chosen to work with model organisms to overcome the methodological barriers inherent in observational studies. I show how their work has generated epistemic progress on central research questions that have not been achieved by those working only with human populations. I close with some observations about how the concept of methodological shortfall can contribute to broader discussions within philosophy of science.

3.2 Philosophers, Fixation, and Methodological Shortfall

3.2.1 The many ways science can operate counterintuitively

A primary area of research for philosophy of science has been the examination of instances where science or scientists have acted counterintuitively. Part of this examination has involved attempts to normalize these patterns of activity. Instead of continuing to view this activity as counterintuitive, philosophers of science (and others) should see these behaviors as normal features of science. On the other hand, philosophers have also described instances of science operating counterintuitively to call out how science is going wrong and (potentially) provide a means of righting it. Three notable philosophers of science who have explored counterintuitive patterns of operation in the sciences are Thomas Kuhn, Imre Lakatos, and Elisabeth Lloyd. Each offers a different characterization of counterintuitive scientific activity. The first two sought to normalize some instances of counterintuitive activity, while the latter offered a normative account that both described how science can go wrong and made suggestions for how scientists can course-correct.

Thomas Kuhn is best known for *The Structure of Scientific Revolutions* (Kuhn 1964). In this work he provides a descriptive account that distinguishes between normal science and revolutionary science. For Kuhn, sciences are shaped by paradigms, which are collections of theoretical claims, methods, and habits scientists have about some domain of natural phenomena. It is the application of these claims, methods, and habits in the solving of problems within this domain that constitutes a paradigm. Throughout history, Kuhn argues, scientific change occurs when one paradigm replaces another, entailing a

revolution in how scientists understand and engage with the world. Scientific work performed within a paradigm is "normal science" and is characterized as well organized. Normal science occurs when the scientific community agrees on the main targets of scientific work, the methods used to explore scientific questions, and how to evaluate the results.

During normal science, researchers are engaged in what Kuhn calls "puzzle solving": attempting to understand the world through the framework provided by the paradigm. This puzzle solving is focused on phenomena in a domain that ought to be understandable through that framework. During this period of puzzle solving, fundamental aspects of the paradigm go unchallenged. Instead, puzzle solving extends the paradigm theoretically and experimentally in minor or gradual ways. However, anomalies can arise. Anomalies are empirical findings that do not fit in the paradigm. We might think that retaining a paradigm that has anomalies is a counterintuitive thing to do. Rather, in the face of anomalies, we would expect scientists to give up or modify their theories. However, as Kuhn notes, scientists don't typically do this. Normal science continues even in the face of mounting anomalies until the number of anomalies accumulates to a point where scientists begin to disagree about the fundamentals of the paradigm. At that point, according to Kuhn, science is in crisis.

Kuhn notes that sciences faced with a critical number of anomalies do not undergo scientific revolution until there is a new candidate paradigm. When a new paradigm is adopted, scientists can again be described as engaging in "normal science." This new paradigm is not directly comparable to the previous one because they do not share common

standards of evaluation (i.e., they are incommensurable). Thus, discontinuity exists between the previous scientific paradigm and the new, post-revolutionary paradigm. Importantly, Kuhn thinks the a period of normal science, followed by a crisis that yields a revolution and a new paradigm, is the regular process for scientific change throughout history. Kuhn's analysis of paradigm persistence in the face of mounting anomalies takes a counterintuitive pattern of activity and normalizes it so that it becomes an expected way for science to operate.

A second example where a philosopher of science analyzed how science can operate counterintuitively is Imre Lakatos and his account of scientific research programs. For Lakatos, the counterintuitive behavior was the persistence of a research program even when researchers failed to make epistemic progress. Lakatos's account of scientific research programs arose as an counter to Kuhn's idea that scientific change was seemingly irrational because paradigms exhibit incommensurability (Lakatos 1970). Lakatos held that scientific research programs were similar to Kuhn's paradigms (i.e., collections of claims, methods, and habits), except that there were generally multiple research programs extant within a given field at the same time. Science proceeds through competition between research programs within a given field or for a domain of natural phenomena. Each research program contains a hard core, or a set of fundamental ideas or commitments, which is "shielded" by a protective belt of less fundamental ideas used to apply these theoretical commitments to the world. For example, 18th century Newtonian mechanics had a core made up of the three laws of motion and a gravitational law. Its protective belt

encompassed ideas about how those laws were applied to particular phenomena (e.g., falling bodies or planetary motion).

For Lakatos, scientific change within a research program occurs only in the protective belt, never in the core, and should typically be progressive. For Lakatos, progressive research programs modify the protective belt such that the theory is extended to new phenomena and thereby increases that theory's predictive power. However, sometimes modifications do not yield these outcomes. A research program is considered degenerating if its modifications of the protective belt fail to extend the reach of the theory or yield empirical progress. Interestingly, the history of science seems to show examples of researchers holding on to a research program even though it is degenerating. This is counterintuitive. One would expect scientists to abandon a degenerating research program. Lakatos argues that it is acceptable for scientists to continue to pursue a degenerating research program, at least for a time. This is because there is a chance that the research programs will eventually return to a progressive form of change by encompassing new phenomena or increasing in predictive power. However, Lakatos provided no guidance on *when* scientists are justified in abandoning a degenerating research program, merely stating that sticking with a degenerating research program is a high-risk strategy (Lakatos 1970).

An example of this counterintuitive pattern of activity can be seen in Urbach's analysis of the environmentalist program in IQ studies (Urbach 1974). Urbach argued that the environmentalist's attempt to explain in-group differences due to external environmental causes was a degenerating research program compared to the hereditarian research program focused on internal causes. However, the environmentalist program

recovered in the mid-1980s with the discovery of the Flynn Effect (Flynn 1987). This refers to the observations made by James Flynn on the rise of standardized IQ scores over time. These observations showed that intelligence, as measured by IQ scores, is not fixed, as hereditarians argued, but rather fluid. This led to the revival of the environmentalist's position. The view we are left with from Lakatos is one in which scientists can sometimes be justified in pursuing a degenerating research program and this pattern of activity can be considered a normal feature of science.

A third example of a philosophical analysis of counterintuitive behavior is found in the work of Elisabeth Lloyd (Lloyd 2015). In contrast to Kuhn or Lakatos, Lloyd was not showing that some counterintuitive activity pattern should be seen as a normal part of science. Instead, she wanted to highlight that science was going wrong in certain domains. Lloyd's work explores the counterintuitive operation of research communities that exhibit a lack of response to increasing criticisms of their theoretical framework. Instead of acknowledging these criticisms and adjusting their research in light of them, some communities continue to pursue their research regardless. She argues that this can occur when the logical structure of their research questions insulates these communities from criticism by determining the kinds of answers deemed appropriate for a particular research question. Since the structure of the question frames what we see as a possible answer, communities can become insulated from criticisms that suggest answers not in the logic of inquiry. Lloyd argues that we should expect research communities to grapple with negative evidence for or criticisms of a theory. Properly working scientific communities exhibit this feature.

Lloyd provides illustrations of this counterintuitive pattern of activity in adaptationist approaches to explaining human and other traits. For example, within studies of the evolution of the female orgasm, researchers often frame their research question as: "What is the function of the female orgasm?" This framing entails that reasonable answers must exhibit the form "the function is *X*." This greatly narrows the possibility space to only those answers with the logical form described above (e.g., an answer like "there is no function" is ruled out). Constraining the set of possible answers has grave consequences, including setting up onerous burdens of proof (e.g., to show that there is no function), or seeing theories as mutually exclusive when they can be compatible or cooperative (e.g., the male orgasm has an evolutionary function, but the female orgasm is a byproduct). Most importantly, there may be no stopping rule (e.g., if you figure out *X* is not the function, you just keep looking at other options, *ad infinitum*). However, Lloyd notes that alternative framings of the research question exist. For example, researchers could frame it as: "Does the female orgasm have a function?" This alternative question framing is still oriented towards understanding the evolution of the female orgasm, but the set of possible replies is broadened to include evolutionary forces other than adaptation, such as developmental constraints (e.g., those arising from homologous genital structures in males and females). Despite the existence of alternative question framings, as often seen in other areas of evolutionary biology, adaptationists retain the narrower question structure, which locks them into a particular form of reasoning that is not epistemically virtuous. For Lloyd, analysis of the structure of research questions can prompt researchers to recognize when

the set of possible answers to research questions is problematically constrained and to suggest alternatives for scientific communities to right their epistemic ship.

What should be clear from our discussion of Kuhn, Lakatos, and Lloyd is that philosophers have identified many different ways that science can operate differently from our expectations. For Kuhn, normal science can persist, even in the face of accumulating anomalies, at least until a critical number of anomalies is reached and a new candidate paradigm emerges. For Lakatos, degenerating research programs can be retained as a kind of high-risk strategy, where there is a chance they can shift to being progressive. For Kuhn and Lakatos, these are aspects of science that may be unexpected at first, but they should come to be seen as normal—we should not be surprised to see these activity patterns in science. For Lloyd, some counterintuitive patterns of activity, like science not addressing negative evidence or theoretical criticisms, is science going astray, and the logic of research questions can account for some cases of this. In each instance, philosophers have provided a way to understand how science can operate counterintuitively. One key takeaway from this discussion is that there are many ways for a science to operate counterintuitively and, at least in some cases, we should not be surprised by this behavior.

3.2.2 Introducing Methodological Shortfall

Like the three philosophers of science in the previous subsection, I also am interested in episodes where science operates in a counterintuitive manner. Specifically, I am interested in instances where scientific communities continue to pursue certain methods even in the face of increasing criticism and a lack of empirical progress. I refer to this pattern of

behavior as "methodological shortfall." Methodological shortfall occurs when scientific communities focus their research programs on certain questions despite available methods being inadequate for addressing these questions.[7] Under conditions of methodological shortfall, existing methods handicap researchers' ability to definitively answer their own questions. In slogan form, their reach exceeds their grasp. At the time, researchers may or may not recognize that they are failing to make epistemic progress because this can only be fully assessed by evaluating empirical progress in the longer term. Therefore, recognizing methodological shortfall typically entails examining a research community over time, often decades.

Methodological shortfall is distinct from other issues that arise with respect to methods, such as when scientists use methods that are *unsuitable* for addressing a question of interest. Research methods are unsuitable when they fail to track relevant features of the world (features that, given existing knowledge, are known to be of interest). For example, using unaided sight is an unsuitable method for learning about cellular organelles. This assessment is based on our existing knowledge that many cells are 30 micrometers in length (the egg cell, at 130 micrometers in length, is the largest cell in the human body and is just barely able to be seen with the naked eye). Whereas the

[7] Methods here refers to "the suite of instruments and techniques that researchers use to produce, record, and process empirical data" (Boyd and Matthiessen 2024,122). For my account to work in practice, one might argue that I need to provide an account of how methods are individuated. A sufficient guide for individuation is that methods can be individuated by differences in the way they engage with the target, the physical instruments and forms of data analysis used, and their associated background knowledge (Boyd and Matthiessen 2024). However, for my purpose, it is sufficient to say that some methods are more like each other than they are compared to other methods and that similar methods can be categorized into groups along these lines.

unaided human eye is only able to discern objects of around 0.1 millimeters, organelles are inside a cell, making them far smaller than what unaided human vision can see. Even someone with the best vision cannot overcome this basic limitation. Therefore, unaided human vision is an unsuitable method for learning about cell organelles given our existing knowledge (i.e., about what is discernible to the human eye and the size of cells) (Alberts et al. 2002).

Although some methods are unsuitable for answering a particular research question, others are suitable but not necessarily sufficient. For example, light microscopy can be used to learn about some cell organelles, like the nucleus. It allows for enhanced discernment at much smaller scales than unaided vision, providing scientists with a window into the cell. This method is suitable in the sense that it can pick out some features that are relevant to the research question at hand given existing knowledge. The light microscope can allow researchers to see objects at much smaller scales, scales which align with the size of some cellular organelles. However, this window is also limited. Light microscopy only allows for discernment of objects of 0.2 micrometers in size. Although researchers can see some organelles using this method, they may not be able to see all of them or any of them at a fine-grained level. Thus, methods can be suitable for addressing a research question, but they may not be sufficient. Light microscopy is not sufficient for learning about the structure of many cellular organelles, such as the Golgi apparatus.

A method is sufficient when it is both suitable and can discriminate between competing hypotheses that bear on the research question at hand. Light microscopy is not sufficient for providing a detailed understanding of the Golgi apparatus because it does not

allow for the discernment of smaller substructures within the organelle. Electron microscopy can provide these details because this technique allows for more fine-grained viewing of objects an order of magnitude smaller (Bechtel 2006). The development of the electron microscope allowed researchers to describe the smaller component parts of the Golgi apparatus, like the small and large vacuoles and flattened cisternal sacs, of which only the latter gave the Golgi apparatus its appearance under light microscopy. Thus, given existing knowledge and the research question, "What is the structure of the Golgi apparatus?", electron microscopy, unlike light microscopy, can provide definitive evidence of Golgi structure.

Importantly, methods that are suitable but not sufficient play an important role in scientific investigations. They provide relevant information or evidence for a research question. Light microscopy provided some evidence that addressed the research question of what the structure of the Golgi apparatus is. This evidence was used as a check to ensure that images from electron microscopy were not artifacts from specimen preparation. In this way, light microscopy still produced evidence that was relevant to the question at hand, but the evidence was not sufficient to fully answer research questions related to the details of Golgi structure. The fact that methods can be suitable and yet insufficient is part of why researchers continue to pursue them. When researchers find methods to be unsuitable for addressing a research question, they give them up. Therefore, the suitability yet possible insufficiency of methods becomes important for recognizing methodological shortfall.

Methodological shortfall is most easily recognized in hindsight. Generally, it becomes apparent once researchers have developed a method to overcome existing

limitations. For example, researchers interested in understanding cellular organelles experienced a period where they became aware that their light microscopy methods were suitable but insufficient. Researchers tinkered with the existing methods by introducing phase contrast and ultraviolet microscopy, which was helpful but still did not overcome the limitation (Bechtel 2006). The degree to which they failed to provide answers to research questions about cellular organelle structure could only be seen after new methods like electron microscopy were developed. This showed researchers what they had been missing—such as the substructures of the Golgi apparatus. Thus, to demonstrate methodological shortfall, one must be able to show that methods are suitable but unable to provide discriminating evidence for possible hypotheses generated in response to a research question. This is easier to do when methods have been developed to overcome these limitations, and a degree of time has passed.

It is worth pausing here to address an objection. What makes methodological shortfall distinct from scientists always needing to refine and improve their methods? Methods may in some sense always be suitable, but they may or may not be sufficient to address a particular research question. It is time scale that distinguishes methodological shortfall from these more benign cases. Methodological shortfall occurs when suitable but insufficient methods continue to be pursued even when there is a lack of epistemic progress for an extended period, often years. The longer these methodological limitations persist, the stronger is the case for methodological shortfall. We can see this clearly in the case of smoking and lung cancer.

Researchers and medical professionals began to see an increase in the number of

cases of lung cancer in the early 20th century. By the 1930s, researchers had begun to investigate if there was a causal relationship between smoking cigarettes and lung cancer. The principal research question was: "What is the causal relationship between smoking and lung cancer in humans?" The first method used to address this question was case-controlled (observational) epidemiology studies.[8] The first of these studies compared 86 lung cancer cases in humans with a similar group of humans who were cancer-free (i.e., controls). This method was able to track the incidence of cancer within two subpopulations: those who smoked and those who did not. By tracking these features, researchers could determine if there was a difference in rates of cancer between these subpopulations—a necessary feature of a causal relationship between smoking and lung cancer. Data from this study showed that people who had lung cancer were more likely— far more likely—to smoke than the non-cancer control subjects. While this method was suitable for producing relevant evidence to answer the research question, it was not sufficient.

The evidence provided by observational methods could not rule out the possibility that the positive correlation between smokers and lung cancer was due to some confounding factor left uncontrolled by the method. For example, the method could not distinguish between: (1) smokers are "a different kind of people" who smoke because they

[8] Although my analysis of the case of smoking and lung cancer uses a distinction between observational and experimental methods to produce relevant evidence, I am not committed to the idea that experimental methods will, in principle, always produce better data than observational studies for a given research question. I am sympathetic to arguments that the epistemic advantages of different research methods are derived from their ability to improve signal clarity, better characterize background factors, and better track and discriminate between the variability of precipitating conditions (Boyd and Matthiessen 2024). It may be the case that utilizing alternative observational methods is sufficient to answer the research question. Seeking experimental methods may not be necessary or may be inappropriate for a given research question.

are “different” and get lung cancer because they are “different,” and (2) smoking causes lung cancer. Subsequent studies were conducted, each attempting to modify the methodology to control for confounding variables. Researchers conducted prospective studies using large study populations in which each smoker was paired with a nonsmoker, controlling for an extraordinarily large set of confounding variables. However, even as researchers tinkered with existing methods over several decades in an attempt to account for confounding factors, some critics argued that the evidence produced by prospective studies was still insufficient to distinguish between the two claims.

Although this debate raged for decades, methods were developed for use in experimental studies of non-human organisms that could distinguish between these two claims. These methods controlled confounding variables so that the only difference between study populations was exposure to smoke and smoke-derived substances. However, these methods were not applicable to human populations for ethical and practical reasons. Thus, an added aspect of relevance for using evidence produced by these methods was their fit with current knowledge about whether the models of study were sufficiently like humans in ways that were relevant to the research question. Most researchers agreed that, given the current state of knowledge, model organisms (mice, in this case) were sufficiently similar to humans (in relevant domains) to use evidence produced by these studies to draw conclusions about humans. Experimental studies in non-human organisms had begun outside of the United States in the 1930s. By 1953, an experimental study in which the backs of mice were painted with cigarette smoke tar had been cited as proving the case against tobacco beyond a shadow of a doubt (Parascandola 2004). It was the

evidence produced by experimental methods in model organisms that was sufficient to distinguish between the two possibilities listed above: (1) smokers are "a different kind of people" who smoke because they are "different" and also get lung cancer because they are "different," and (2) smoking causes lung cancer. Evidence from experimental studies showed that smoking causes lung cancer.

In this case, researchers faced a prolonged period in which they continued to pursue suitable methods, observational prospective studies, even though they were unable to fully distinguish between the two claims. Tinkering with the method did improve the quality of the evidence, for example by detecting a larger effect size and controlling for confounding variables. Yet this methodological tinkering did not remove existing limitations. It was only by leveraging evidence produced by non-human experimental methods—methods which did not exhibit the same limitations—that researchers could finally make epistemic progress and say with sufficient certainty that smoking caused lung cancer.

Recognizing instances of methodological shortfall (i.e., suitable but insufficient methods for answering a research question) can help researchers begin to make epistemic progress because those facing it are left with two choices: (1) work with available methods to try to overcome the existing shortfall (such as by elaborating on existing methods), or (2) adopt different methodological approaches. Adopting different methodological approaches may be as simple as leveraging the evidence produced by existing research from other research communities that already utilize different methods (as we saw in the smoking and lung cancer example). In the absence of this existing research, the second option often involves developing or learning new methods and redirecting resources.

Scientists have already spent a lot of time and energy learning how to work with their existing methodologies. Switching to a new method has a high upfront cost. What might motivate researchers to choose option two, especially given the cost? The first answer is frustration with not making adequate empirical progress. However, frustration alone does not necessarily yield a compelling justification. I argue that scientists can be logically justified in making this switch in circumstances where the adoption of different methodology has demonstrated the potential to yield results that are conducive to empirical progress.

As we saw with smoking and lung cancer, researchers facing methodological shortfall first engaged in tinkering. When critics claimed that prospective studies could not yield causal information because they did not rule out reasonable confounding variables, researchers tinkered with existing methods to extend their empirical reach by performing larger studies that controlled for more potential confounders. However, in this case, even the extensions of these methods of research were not enough for researchers to secure an unambiguous answer to their question.

One feature of methodological tinkering that we might expect to see, but is largely absent from the above example, is a kind of technological optimism. This may occur when researchers acknowledge that there is a gap between their methods and their aims, but express confidence that a new technological development is just around the corner and will solve the issue. When new technologies are developed, researchers often jump to utilize them. However, these can still harbor similar limitations. Yet the consistent availability of new technologies can breed an optimistic attitude that tends to downplay methodological

shortfall. This is because, in the short term, new technologies can appear conducive to progress, but time reveals that these methods suffer from the same limitations they were meant to overcome. Like Lakatos's account of degenerating research programs, there is always a possibility that new technologies and method tinkering might extend a field's reach enough to grasp a sufficient answer. However, as methodological shortfall persists, there ought to be increasing pressure to develop or leverage new methods to bear on a research question. Failing to do so amounts to science going wrong.

In the following section, I use contemporary behavioral genetics as a case study to demonstrate how the concept of methodological shortfall can help us make sense of ongoing issues with contemporary behavioral genetics research performed in humans. Analyzing these issues in behavioral genetics through the lens of methodological shortfall reveals two paths forward for human behavioral genetics research: (1) continue to use increasingly sophisticated observational studies like GWAS and Twin Studies or (2) adopt experimental methodologies that can only be used in non-human model organisms, thereby approaching their object of study in an indirect way.

3.3 Methodological Shortfall & Behavioral Genetics

3.3.1 Behavioral Genetics: Research Questions & Methods

To demonstrate how methodological shortfall is exhibited by human behavioral genetics,[9] I first introduce a set of related questions that have shaped and continue to shape the field

[9] Importantly, behavioral genetics is a field that encompasses research performed in both human populations and model organisms. At this point, I focus on research questions within human behavioral genetics.

as well as motivate the methods available to answer these questions. I do this to establish that the methods employed by the field are suitable for pursuing the field's research questions. To provide a basic picture of the research questions within behavioral genetics (BG), I rely on the prior philosophical work of Helen Longino (Longino 2013b).

BG is concerned with understanding how genetic and environmental variation relates to variation in particular behaviors. One of the central aims of this research is to quantify how much of the behavioral variation is due to genetic variation versus environmental variation. There is consensus among BG researchers that both genes and environment contribute to variations in behavior. Current research focuses on understanding the relationship between genes and environment and their respective influences on behavioral variation. Researchers approach these questions at the population level (rather than in individuals), by taking stock of genetic and environmental variation within a population and showing how it contributes to behavioral variation manifested within the same population.

Not all researchers are satisfied with providing an account of the relationship between genes and environment in determining behavioral variation at the population level. Some view behavioral genetics as a means to uncover relationships between genetic variation and behavioral variation at the level of the individual. This is present in the claim by many proponents of human BG that this research can help uncover causes and mechanisms underlying behavior. This kind of reasoning is often used to motivate studies. For example, in a recent GWAS study of educational attainment, researchers stated that a potential payoff from their study was that its results could be used to "disentangle the

mechanisms by which genetic factors affect EA [educational attainment] and cognitive phenotypes"(Lee et al. 2018,11). A similar appeal to mechanisms is also evident in BG studies using Twin and Family Study approaches. For example, in a recent Twin Studies review article the authors claim, "They can also extend the classical twin design to model causal relations between exposures and traits" (Hagenbeek et al. 2023, 849). In addition to the possibility of gleaning information about genetic causes of differences in individual behavior, some argue that BG can indicate the limits of certain environmental interventions on behavior and the populations most likely to benefit from these interventions (Burt 2024; Harden 2021a; K. Paige Harden and Koellinger 2020). Although there is some heterogeneity of work withinBG, the field's main research question can be summed up in a slogan: "Why do individuals behave as they do?" (Longino 2013,4).

Despite this shared orientation, it is worth noting variation in how this question is instantiated within particular research projects in the field. As Longino has previously demonstrated, this slogan can appear as: "What does genetic variation within a population contribute to behavioral variation?"(Longino 2013,4). It can also be subdivided into finer grained questions that are shaped by existing methods. While a complete survey of these is outside the scope of this paper, a brief overview of their diversity is useful. Fine-grained research questions can include (but are not limited to):

- To what degree is behavior heritable? (How much are differences in parents correlated with differences in offspring? How much do differences in parents influence differences in offspring?)

- How much of the difference in expression of behavior in a population is associated with genetic differences?
- Does the heritability statistic, and hence, the purported degree of genetic influence on behavior, change over time (both with age and over generations)?
- How can the methodologies used to study genetic influences on behavior be refined and extended?
- What genomic regions or genes are associated with behavioral variation?
- How do regions or genes contribute to behavioral variation?
- What is an individual's probability of exhibiting some form of behavior that varies in the population?

With a better understanding of its research questions, it is now possible to assess whether the methods used in BG research are suitable for answering them. Human behavioral geneticists rely on a subset of quantitative genetics methods. Experimental versions of these methods are not used because they would be both deeply impractical and highly unethical. Experimental methods involve researchers tightly controlling both the organism's environment and genetics through careful breeding and standardized living conditions. This makes it possible to break complex correlations between different variables. Human BG researchers only work with observational quantitative genetics methods. There are four main kinds: twin and family studies, candidate gene studies, GWAS, and Polygenic Score (PS) studies.

Classical Twin studies operate by measuring a trait of interest in both monozygotic (MZ) and dizygotic twins (DZ). MZ twins share the same genome, whereas DZ twins share 50% of their genome. Researchers also assess shared and non-shared environments. By measuring a particular trait in the population of MZ and DZ twins, researchers can assess the total variance of a trait due to additive genetic effects, dominant genetic effects, shared environment, and non-shared environment (Posthuma et al. 2003). Twin Studies have been a longstanding method in behavioral genetic studies (Boomsma, Busjahn, and Peltonen 2002). Using this method, investigators obtain a heritability estimate, which is the percentage of the total variance of a phenotype due to additive genetic effects. Importantly, the heritability measure applies only to the population used to generate it and does not apply to individuals. This method is suitable for addressing at least a subset of the field's research questions because the method is able to track relevant features, such as genetic variation (via DZ and MZ twins) and environmental variation (via shared and non-shared environment). Therefore, this method can produce evidence that speaks to the relationship between genetic, environment, and behavioral variation in the form of heritability estimates.

Researchers can leverage these heritability estimates to test for gene x environment interactions for a given trait by including environmental measurements to stratify the twin samples. For example, in addition to measuring happiness for a set of MZ and DZ twins, researchers can also include a binary variable for being married or unmarried. If the heritability estimate differs between the two environments (married/unmarried), it provides evidence for gene x environment interactions. The classical Twin Study design can be

extended to include parents and also to study cultural transmission and gene x environment co-variance. It can also be altered to include the offspring of twins to study gene x environment correlations (Hagenbeek et al. 2023). Although twin studies allow for an estimate of the genetic contribution to behavioral variation, they cannot be used (on their own) to identify specific regions of the genome that underlie this genetic contribution.

The candidate gene approach was one of the first observational quantitative genetic methods used in humans that linked behaviors to specific genomic regions or genes. The candidate gene approach relies on prior knowledge about an established or hypothesized relationship between some behavior and a gene. For example, using a candidate gene approach to investigate how genetic variation relates to variation in the severity of depression might rely on existing hypotheses about the mechanism of depression, focusing their study on a promotor region of a particular serotonin transporter gene. Researchers would then establish sequence variants (polymorphisms) for a given gene or genomic region, as well as characterizing the variation in their behavior of interest. This characterization could be binary, such as presence or absence of a behavior (a common approach in behavioral pathologies), or involve a range of variation. Researchers then test a given sample of individuals for an association between the behavioral variation and genetic variation (i.e., for the specific gene or small genomic region). The results may indicate an association between the sequence variation of the candidate gene and the behavioral variation. This approach was popular in the early days of genomic sequencing, when even sequencing a single gene for a study sample was quite laborious and expensive. This method was clearly suitable when it was first used; it tracked genetic variation

hypothesized to be relevant and behavioral variation. However, continued research showed that many (i.e., thousands of) genes rather than a few might contribute to behavioral variation. As a result, this method became less suitable. Given our current state of knowledge, the candidate gene approach is unsuitable. Researchers have since abandoned it. As the cost of genetic sequencing decreased, researchers adopted approaches that examined genetic variation at thousands of points along the genome, rather than at a single gene.

The Genome Wide Association Study (GWAS) is one such approach. GWAS works by gathering data on genomic and phenotypic (in this case, behavioral) variation from a population of individuals. Researchers start by comparing thousands of points along a genome, single nucleotide polymorphisms (SNPs), to those same points in each genome sampled; the genomes sampled make up the study population. Then, researchers compare the differences between the phenotype of interest—how individuals differ in a particular attribute—to differences in their genomes (i.e., differences at specific SNPs). This is done for traits measured as either present or absent, such as having schizophrenia or not having schizophrenia, and traits measured as continuous variables, like height. In addition to the standard GWAS methods described above, there are also within-family models. They work by adding additional information to statistical modeling, like the deviation of an individual's genotype from the mean sibling genotype(Howe et al. 2022)[10]. These models attempt to control for potential confounding factors such as between-family contributions

[10] For example, in a sibling pair where one sibling has two risk alleles and the other sibling has one risk allele, the mean sibship genotype is 1.5 risk alleles and the individual's deviations are +0.5 and −0.5, respectively (Howe et al. 2022).

of SNPs. Estimations of direct genetic effects generated from these designs are thought to be less biased, especially for traits affected by demographic variables and indirect genetic effects.

A GWAS generates several different kinds of information. First, it can yield an estimate of the variation in a trait that can be explained by genetic variation. This estimate is generally reported as a percentage and applies to the population sampled, not the individual. Second, a GWAS can also be used to identify quantitative trait loci (QTL), which are points on the genome significantly associated with a particular trait. A GWAS also generates estimates of the sizes of SNP effects, which are understood to be the marginal effects of having particular SNPs. In addition to these estimates, GWAS results can also produce a PS (i.e., a polygenic score or a polygenic risk score). Through GWAS, researchers identify SNPs with significant associations to a trait of interest; then they can generate PS. A PS is calculated by taking the sum of QTLs and weighting them by their effect size (Choi, Mak, and O'Reilly 2020). A PS provides a number that is a measure of the additive genetic propensity for some behavior of interest. PSs have been interpreted as genetic predictors of behaviors in a population but have poor individual predictive accuracy for behavioral traits.

Overall, a GWAS yields the following: Information about the population regarding the percentage of phenotypic variation explained by genetic variation, regions of the genome within a population that are associated with a particular trait (and the effect of having different SNPs at those locations), and a predictive measure of the additive genetic propensity for a given behavior. This method is clearly suitable for addressing at least a

subset of the field's research questions because the method can track relevant features of the world, given existing knowledge. These relevant features include genetic variation (cashed out here as common single nucleotide polymorphisms) and behavioral variation (difference in the values of behavioral parameters). This method can also assess associations with some environmental parameters, such as income, family status, or nationality (among others) if this information is available for individuals whose sequences were collected. Therefore, this method can produce evidence that speaks to the relationship between genetic, environment, and behavioral variation; it does so by assessing how much of the variation in behavior is associated with genetic variation and using the results to predict behavioral measures.

Methods employed by human behavioral geneticists are suitable; they can capture relevant features of the world given our current knowledge. They allow the tracking of genetic, environmental, and behavioral variation, although the way these methods track these features differs. However, whether these methods are sufficient remains to be seen.

3.3.2 The Mounting Evidence of Methodological Shortfall

With both BG research questions and methods on the table, it is time to assess the evidence for methodological shortfall. Two features of methodological shortfall should be kept in mind. First, the defining feature of methodological shortfall is that researchers continue to use suitable methods even though these methods are insufficient, that is they fail to discriminate between possible hypotheses. Insufficient methods are limited in that they cannot provide evidence of the kind needed to address research questions with appropriate

confidence. Second, this limitation must persist over time, often on the order of decades. Sometimes when methodological shortfall extends over this timescale, we can see a kind of perennial technological optimism. Researchers continually adopt new technologies in the hope that these will extend the reach of their methods. This appears frequently in human behavioral genetics. Scientists working with available methods are often aware of their limitations, the way in which their methodological reach exceeds their grasp. However, there is a persistent appeal to new technologies to bring their aim within reach.

The first classical Twin Study was published in the early 1920s.[11] Since these initial studies, Twin and Family Studies have become the major workhorse of human BG. However, there are existing limitations to these methodologies. A key limitation of Twin and Family Studies is that they assume MZ and DZ twins share equally similar environments. Without this assumption, much of the ability to distinguish between genetic and environmental effects becomes confounded (Friedman, Banich, and Keller 2021). There is growing evidence that this assumption is often violated, which in turn has led some critics to argue that Twin and Family Studies overestimate the heritability of certain traits (Harden 2021a; Charney 2012a). This overestimation stems from the confounding of genetic heritability with the inherited environment.

Another assumption present in most Twin Studies is that genetic and environmental contributions are independent. Researchers acknowledge that this assumption may be

[11] There is some debate regarding which study is the first classical Twin Study. Some argue it should be two papers published in 1924 one by Siemens and one by Merriman, while others argue an earlier paper by Walter Jablonski in 1922 should claim the title.

violated, especially for behavioral traits (Friedman, Banich, and Keller 2021). This results in gene-environment correlations. Independence can be violated in several ways. First, when the same genetic factors are passed down to one's offspring (where they impact behavior and are referred to as direct genetic effects), they also can operate within the parents to shape the offspring's environment. This is referred to as passive gene-environment correlation (rGE) in Twin Studies literature and genetic nurture in the GWAS literature. Active rGE occurs when individuals select environments that better fit them as they age (Willoughby, Polderman, and Boutwell 2023). Evocative rGE occurs when an individual's traits that exhibit genetic contributions evoke responses from the environment and other group members (Willoughby, Polderman, and Boutwell 2023). Twin and Family studies also assume that trait variance is due to additive genetic effects, which assumes little to no genetic correlation or interaction, as well as no gene-environment interactions.[12] Given these violations of the underlying assumptions upon which this methodology is based, the extent to which researchers using only Twin or Family Study data are able to distinguish between direct and indirect genetic effects is unclear at best. Researchers have been aware of these limitations for over two decades and are still trying to overcome them (Jaffee and Price 2007). They have attempted to leverage data from GWAS studies to discriminate between direct and indirect genetic effects, but this method also faces limitations (see below). In short, Twin and Family Studies alone are insufficient to discriminate between competing hypotheses generated in response to the research

[12] Gene-environment interactions are distinct from gene-environment correlations. The former occurs "when the magnitude of genetic effects are dependent on, and differ across, varying aspects of a given environment"(Willoughby, Polderman, and Boutwell, 2023,3)

questions of interest (e.g., How much of the difference in expression of behavior in a population is associated with genetic differences?).

As we saw above, the candidate gene approach could not fulfill its promise of finding a small subset of genes that would explain behavioral variation (Tolan and Leventhal 2017).[13] This was partially due to the fact that the vast majority of these studies were underpowered, and many reported results could not be replicated. Additionally, candidate gene research is most applicable when only a handful of genetic polymorphisms are responsible for the variation in a particular trait. With the advent of GWAS, it became evident that hundreds of genes underlie most behavioral traits. As relevant knowledge was updated, the candidate gene approach shifted from being suitable to being unsuitable and was subsequently abandoned.

GWAS came on the scene in the mid-1990s but, at first, the cost of sequencing made it prohibitively expensive. However, as the cost of genetic sequencing began to fall, researchers flocked to this method. In 2005, the first GWAS that sought to identify SNPs associated with macular degeneration was published. This methodology was adapted for the study of human behavior soon after. Twenty years later, we know much more about the limitations of GWAS for the study of human behavior. The first is that GWAS does not locate causal variants. GWAS only examines a subset of SNPs that are preselected. These SNPs are used to identify quantitative trait loci (QTLs). However, these QTLs "can be

[13] There are further issues with candidate gene approaches that I ignore here. For example, they rely on existing knowledge or hypotheses about the mechanisms of how genes and behavior are connected, but this functional information about genes is often missing or, at best, extremely limited.

large (>1Mbp) and contain dozens of genes, hundreds of SNPs, and several hundred thousand variants" (Burt 2024, 9). While attempts have been made to narrow down QTLs to more specific causal variants, limited information about gene function hinders this project (Schaid, Chen, and Larson 2018; Nicholls et al. 2020; C. H. Burt 2024). What GWAS can provide is a measure of direct genetic effects. It is meant to provide insight into biological mechanisms, or the "causal path" from genotype to phenotype (C. H. Burt 2024). However, the accuracy of these estimates relies on the assumption that variants must be randomized across genetic and trait-relevant environmental backgrounds (Veller and Coop 2024; Weir 2008). This assumption is often violated. As a result, GWAS estimates (and resulting PSs) are confounded by population stratification, assortative mating, and associative effects (Abdellaoui and Verweij 2021; Lawson et al. 2020; Veller and Coop 2024;Young et al. 2019).[14] This is especially true for complex social traits (Burt 2024). Not all studies that utilize GWAS data aim to discriminate between direct genetic effects, or between genetic and environmental effects. Therefore, the limitations imposed by these confounding factors are most relevant to research questions related to direct genetic effects or studies where discrimination between genetic versus environmental effects is at issue (Burt 2024). However, GWAS data typically plays this role in social science research. Thus, we should be concerned that GWAS faces a notable limitation; it is confounded in its ability to discriminate between genetic and environmental effects, which causes direct

[14] Population stratification is the presence of a systematic difference in allele frequencies between subpopulations. Assortative mating is when individuals preferentially choose to mate with those more similar to themselves (rather than mating randomly). Associative effects are when the phenotype of an individual is influenced by the genetically influenced phenotypes of others they are strongly related to (C. H. Burt 2024).

genetic effects and genetic influences to be overestimated. In other words, the method is suitable but not sufficient because it cannot discriminate between different hypotheses proposed in response to central research questions.

I have now shown how current observational quantitative genetic methods in human BG fail to discriminate between genetic and environmental influences on behavior, severely curtailing their ability to identify causal effects that would answer research questions directed towards establishing how much of the differences in expressions of behavior in a population are associated with genetic differences. These limitations, which have persisted for more than a decade, prevent the achievement of research aims concerned with understanding why individuals behave the way they do. In other words, these methods are insufficient to discriminate between genetic and environmental influences; these methods fail to control confounding variables and could only do so via experimental interventions that are prohibited for human subjects. Researchers are aware of these limitations and are engaged in tinkering with existing observational quantitative genetic methods to overcome them. As we will see, this tinkering has only recapitulated the same limitations it sought to remove. Yet, researchers continue to insist that these methods provide the most promising way to address their research questions. This suggests that human behavioral genetics exhibits methodological shortfall.

3.4 Two Ways Forward

When faced with methodological shortfall, scientists have two paths forward: (1) tinker with existing methods, working to extend their reach, or (2) adopt new methods that are

better positioned to achieve their goals. Historically, human behavioral geneticists have favored the first option, even in the face of mounting criticism.

3.4.1 Tinkering with the Method

In the case of classical GWAS, researchers have attempted to mitigate the effects of confounding variables by conducting studies only within populations of a shared ancestral group, normally white Europeans. Doing so is thought to reduce between-group population stratification. However, population stratification does persist even within ancestral groups. There have been several statistical methods used to reduce within-group population stratification. For our purposes, the details do not matter. It is sufficient to say that these techniques "*reduce but do not eliminate* relatedness confounding" (Burt 2024, 13, emphasis in original). For example, a common adjustment used in GWAS studies is linkage-disequilibrium score regression (LDSC).[15] This technique is used to separate population stratification biases (i.e., spurious correlations) from causal genetic effects. However, this method relies on certain assumptions that are expected to be violated by population stratification, such as no long-range linkage disequilibrium. In addition, there are other effects that are not controlled by this method. Thus, even a prevalent technique used to control for confounding factors is itself subject to assumptions that are likely violated (Border et al., 2022; Bulik-Sullivan et al., 2015). At best, it is unclear if this technique (and others like it) can adequately mitigate confounding factors. At worst,

[15] Linkage disequilibrium concerns the frequency of association between different alleles being higher or lower than expected by chance. This nonrandom co-occurrence of genomic segments in a population is a form of stratification bias.

estimates of heritability, meant to yield SNP effect sizes, include genetic and environmental confounds that inflate the effect size (Burt 2024).[16]

Researchers have also performed within-family GWAS, which is thought to control for shared genetic and environmental backgrounds. This methodological improvement works by estimating sibling differences in a trait, based on sibling differences in genotype, controlling for mean genotype. Comparing results with this method to conventional GWAS shows that conventional GWAS estimates are confounded. These confounds can be mitigated (but not eliminated) by within-family GWAS. Confounding is especially present for complex social traits (Howe et al. 2022). Within-family GWAS estimates for direct genetic effects are lower by one half or two thirds compared to conventional GWAS estimates (Howe et al. 2022; Lee et al. 2018; Brumpton et al. 2020; Mostafavi et al. 2020). However, within-family GWAS does not eliminate confounding due to indirect effects, such as gene interactions or pleiotropy (GxG) or gene-environment interactions (GxE). Yet there is an even more glaring issue with this improved approach. Although it reduces confounding factors, it can inadvertently change the research question in a subtle way. That is, the question about the causes of differences between siblings in the same family may be different from the question about the causes of differences between unrelated individuals in a population (Burt 2024; Coop and Przeworski 2022a). Put another way, there is an additional question about whether findings from Twin and Family Studies generalize to unrelated populations. What makes family members different from each other could be

[16] SNP effect size refers to the contribution of a SNP to variation in the trait of interest.

different from what makes unrelated individuals different from each other. Thus, while this method reduces the confounding factors present in conventional GWAS, it may not provide information relevant to the wider question that conventional GWAS attempted to answer.

In terms of Twin and Family Study methodologies, researchers have focused on adapting standard methods to account for GxE and rGE. Although there has been much theoretical work on GxE, there has been little empirical work (Hagenbeek et al. 2023). However, rGE has been extensively studied by adding an additional environmental or genetic variable to studies. For example, if researchers are interested in depression, they may add the variable of marital status (a trait-relevant environmental variable) and a PS for depression to their study. If genetic factors influence both depression and marital status, then there is evidence of rGE. This is simply one of many methods used to illuminate rGE. To get at GxE, researchers can test to see if heritability measures change across different environments. If estimates do change across environments, then this is taken as evidence of GxE. Note that the techniques used to characterize violations in methodological assumptions employ the very same methods at issue.

Although previous methodological tinkering has attempted to control or explore confounding present in existing methods, researchers have also attempted to combine GWAS and Twin Study designs. This can be seen in the development of Mendelian Randomization (MR) and direction of causation (DOC) twin models to study causality. MR uses SNP or PS thought to be correlated with the exposure variable but not with confounding variables or the outcome itself. However, this method is generally

underpowered and susceptible to GxG interactions. DOC works by analyzing data in terms of two variables, which are thought to be inherited through different modes, such as shared environment and genes, to determine whether one variable causes another. Combining these two approaches reduces (but does not eliminate) the limitations of each approach. Researchers acknowledge that each approach can get closer to a causal result but that each approach does not escape the limitations that prevent it from unambiguously answering questions central to BG research.

Interestingly, some researchers are adamant that these observational quantitative genetic methods can yield information about biological mechanisms and causal pathways. "Amidst all of this progress, the capacity to more clearly illuminate causal pathways for health and wellness outcomes ranks high among what behavioral genetics has to offer the scientific enterprise" (Willoughby, Polderman, and Boutwell 2023). Yet in the very same article, the authors reiterate the limitations of existing methodologies, namely that causal claims cannot be made. The same authors who tout these methods as promising for revealing causal claims are simultaneously intimately aware of their inherent limitations. For example, they quote a recent systematic review on Twin Studies.

> By controlling for common causes, twin studies can test what have been referred to as 'quasi-causal hypotheses'. The use of phrases like 'quasi-causality' and 'causal inference' (as opposed to straightforward 'causality') highlights that twin studies (like other epidemiological approaches) cannot typically demonstrate causality between two variables—they cannot prove that X (the exposure) causes Y (the outcome)' (Willoughby, Polderman, and Boutwell 2023, 12).

In the face of this situation, Willoughby and colleagues claim that, "if the data are being asked to offer an improvement in causal inference capacity beyond observational research,

they perform well" (Willoughby, Polderman, and Boutwell 2023, 12). However, improvement in methodological reach does not mean these methods have overcome their methodological shortfall. They have merely reduced the breadth of the gap. As the researchers themselves put it: "If twin studies are being asked to offer up the clearest evidence of causation possible, then they are being asked to do something beyond their capacity" (Willoughby, Polderman, and Boutwell 2023, 12).

When faced with methodological shortfall, human BG researchers have largely engaged in methodological tinkering. Although this tinkering has augmented the ability to discriminate between competing hypotheses relevant to their research questions, the core limitations are still present. What would happen if behavioral geneticists supplemented their current observational methods with new experimental ones?

3.4.2 Experimental approaches in model organisms

The observational methods available in human BG are insufficient, however sophisticated they have become over time. They fail to provide discriminating tests between possible hypotheses relevant to the research questions of interest to behavioral geneticists. Methodological shortfall is particularly acute for research questions concerning causes, mechanisms, or biological pathways. Answers to these questions are one of the central promises of current human BG research. As we have seen, one reaction to this methodological shortfall is to tinker with existing observational methods to extend their reach. This is the path many behavioral genetics researchers choose to pursue. However, there is an alternative. Human behavioral geneticists could adopt (or leverage data from) other methods, such as experimental quantitative genetics. These methods are better suited

to the investigation of causal effects, mechanisms, and biological pathways. To adopt these methods, researchers face a significant hurdle. Using experimental quantitative genetics in humans is ethically impermissible because, in part, they require the ability to control mate choice and environment. At present, adopting experimental quantitative genetic methods means learning to work in systems where these kinds of techniques are ethically permissible: non-human model organisms. Alternatively, researchers working in humans could also leverage information produced by these methods for application to their work in humans. This type of information sharing is common in other subdisciplines of genetics, like molecular, functional, and medical genetics. There are BG researchers who do use experimental quantitative genetic techniques in non-human model organism systems. The results from their studies provide a window into what could be learned if human behavioral geneticists opted to leverage these methods.

What does it mean to leverage studies from non-human model organisms? Model organisms are a small subset of organisms commonly used in biological research focused on genetics, development, and physiology. Famous examples include *Drosophila* (fruit flies), *M. musculus* (mice), *C. elegans* (soil roundworms), and *Arabidopsis* (thale cress). Due to their long history of use in scientific studies, shared resources are available to researchers interesting in working with these systems, including breeding stocks, genetic resources, and standardized techniques (Ankeny and Leonelli 2011). These shared resources and history of use in experimental biology are important assets when employing experimental quantitative genetic methods in model organisms.

In Chapter 1, I distinguished between using model organisms as surrogates versus exemplars. When model organisms act as surrogates, researchers gain information about a specific biological system (the model organism) in order to apply this information to a specific target (typically, humans). In contrast, when model organisms are used as exemplars, researchers take the model organism as standing in for a general class of organisms (e.g., vertebrates) and use information from these models to make generalizations to the pertinent group of organisms. Distinguishing between these two uses can be difficult, in part because sometimes both are operative. However, for our discussion, both uses of model organisms are relevant. BG research in model organisms can tell us about genetic and environmental mechanisms underlying behavioral variation generally (as exemplars) and specifically for humans (as surrogates). Adopting these methods permits researchers to overcome existing limitations of observational methods to gain causal insights previously out of reach. To demonstrate this point, I focus on experimental methods in model organisms that are used to explore the equal environment assumption and prevalence of GxG interactions as they relate to our understanding of the genetics of behavior. I show how methods available in model organisms research are sufficient, that is, they can perform discriminating tests between competing hypotheses that are relevant to research questions. These results also illuminate relevant features of the world that methods available to human researchers cannot track, which implies that some methods available in human research are not just insufficient but unsuitable.

Researchers studying humans and those studying model organisms both see genes and environment as causes of behavioral variation. Their current research questions focus

on investigating the complex interactions between genes and environments, especially as they relate to causal mechanisms and biological pathways. However, the two communities differ in the level of complexity of these interactions that can be studied. This difference is driven by methodological shortfall. The methods of human BG researchers have failed to yield definitive tests capable of distinguishing between competing hypotheses providing solutions to their questions about the role of the equal environment assumption and GxG interactions. Their methods simply cannot control confounding variables or reach the sample sizes needed to detect the miniscule effect sizes. As a result, researchers working in humans have remained agnostic about the presence of GxG interactions and the potential impact of violations of equal environment assumptions with respect to their contribution toward individual variation in behavior. This is deeply troubling because understanding the presence and extent of GxG and GxE interactions is critical for characterizing the limitations of existing methods.

Studies in non-human model organisms suggest that violations of equal-environment assumptions are common and serious. Twin Study methodology assumes that MZ and DZ twins have environments that do not differ in ways that matter for the behaviors of interest. That is, the results from Twin Studies are assumed not to be confounded by environmental influences from shared environment, especially in studies of twin reared apart. These studies concentrate on the effect of MZ twins who were raised separately and on how non-shared environment affected behavior. Notably, these studies often include twins that were separated late into childhood (sometimes around 8 years of age) and this is assumed not to affect the study design (Charney 2012b).

In contrast, there are similar but distinct methods from quantitative BG, like experimental cross fostering studies, that are used in model organism research. In these studies, the offspring of one strain of mouse are raised by an adoptive mother of a different strain (i.e., a different genetic background) and then their behavioral variation is assessed. The results are then compared to the behavior of offspring reared by the biological mother. These types of studies attempt to assess the impact of maternal effects on offspring behavior. This study methodology relies on excruciatingly diligent control of the environment, a condition not achieved in Twin Studies. This type of control is necessary to detect any environmental effects that are not due to the mother's influence on the offspring's behavior. Even so, literature on experimental cross fostering is plagued by "cage effects," which are effects attributed to differences in the environment, not maternal influence (Charney 2012b). The persistence of environmental confounds, even within carefully controlled experimental designs, provides further evidence that the equal-environment assumption present in Twin Study methodologies is violated. This violation in turn leads to artificially inflated measures of "genetic" heritability, thereby yielding misleading answers to central research questions like: "How much of the difference in expression of behavior in a population is associated with genetic differences?"

Heritability measures are further compromised when comparing the heritability of behaviors measured in humans with those measured in model organisms. In Twin Studies, the heritability of aggressive behavior in humans has been placed at roughly 50% (Charney 2012b). Studies of aggressive behavior in fruit flies yield a heritability measure of 10% (Edwards and Mackay 2009). Furthermore, when attempting to understand what accounted

for the lower heritability in fruit flies, researchers found that the low heritability was due to high environmental variance. This is surprising given that all flies were reared in the same controlled environment (Edwards and Mackay 2009). Findings for BG studies performed in model organisms suggest that the estimates of heritability present in human studies are confounded, leading to artificially high estimates. While this limitation is noted by researchers, the degree of inflation due to environmental confounds is difficult to assess. Model organism research shows that even when utilizing randomized experimental designs and attempting to directly control variation environment (control that is absent in human studies), effects of environmental variation persist. This suggests that current observational methods are unable to distinguish between genetic and environmental variation in heritability estimates and other results from Twin Studies that rely on this distinction.

As noted in the previous section, most results from BG that claim to deliver estimates of additive genetic effects are confounded by GxG interactions, otherwise known as epistasis. Methods available in human BG are not well positioned to assess epistatic effects. Therefore, there is little evidence of these effects on human behavioral traits. However, research performed in model organisms provides evidence that epistatic effects are present and have a meaningful effect on results. Directly assessing epistatic effects is possible in model organisms like yeast, *C. elegans*, and *Drosophila* due to the large collections of strains that differ genetically only at the point of interest. Researchers also have measures of the trait of interest for these strains. Two-way analysis of variance is used to estimate the main effect of genotype 1, genotype 2, and genotype 1x2 interactions. If the interaction term is significant, then there is evidence of interaction. These results can be

used to estimate not just additive genetic effects but also dominance and epistatic effects on a trait. Importantly, measures of epistatic effects in model organisms are not the same as epistatic variance measures. Measures of epistasis might be produced using Twin Studies, but epistatic effects cannot be inferred from epistatic variance measures. They can, however, be directly measured using an experimental quantitative genetic technique, quantitative trait loci mapping, in model organisms. This method has been used to detect epistatic effects on aggressive behavior in *Drosophila* (Edwards and Mackay 2009).

Detecting epistatic effects is especially relevant to research questions involving the prediction of phenotypes based on genetic information. Models in humans like PSs generally assume only additive genetic effects are present. However, studies in model organism have shown that these models can have low predictive accuracy when there is widespread epistasis. Although techniques in human BG are ill suited to the estimation of epistatic effects, methods present in model organism research can and have done so. The inability to assess epistasis in humans will hinder predictive measures in cases where there is a widespread effect due to genetic interactions. However, researchers have little ability to identify these cases or correct their predictive models in response. Thus, techniques available in non-human model organism inquiry can provide an avenue for gaining traction on epistatic effects currently lacking (and underappreciated) in human BG research.

The example above shows that experimental quantitative genetic methods used in model organisms do not face methodological shortfall. These methods are both suitable, able to track relevant features of the world given current knowledge, and sufficient—able to provide discriminating tests between competing hypotheses pertinent to research

questions. These methods can track features of the world relevant to uncovering epistatic interactions by controlling background variation and environmental features. This ability allows these methods to characterize epistatic effects and distinguish them from additive genetic effects (and other related effects). Experimental quantitative genetic methods in model organisms provide better tracking of relevant features of the environment and reveal violations of the equal environment assumption. However, for researchers working in human BG there can be more to learn than simply that experimental quantitative genetic methods are suitable and sufficient.

What does an examination of the experimental quantitative genetic methods available in model organism research reveal about the observational quantitative genetic methods used in human research? Taken together, these findings suggest that not only are the methods available in model organisms suitable and sufficient for some research questions, but that the methods available to human researchers are ultimately unsuitable. Studies from model organism research reveal aspects of the world that are relevant to the research questions in view and that were previously not known from human research. This demonstrates that methods available to human researchers are unable to track some of the relevant features of the world, given knowledge of these features provided by model organism research. This is most evident from our example of epistasis. Epistasis is relevant in the estimation of direct genetic effects and in predictive estimates of behavior due to direct genetic effects. These estimates, especially predictive estimates, are confounded when there is widespread epistasis. Yet, human researchers lack methods to detect

epistasis. Consequently, they are unable to track features of the world that are relevant to their research questions, which involve estimating direct genetic effects.

Experimental quantitative genetic methods provide answers to research questions. In doing so, they provide the opportunity to update the understanding of human BG researchers with respect to the relationship between genetic variation and environmental variation. Having an updated understanding of the phenomena of interest changes researchers' understanding of the relevant features of the world, the features suitable methods need to track, such as those related to epistatic interactions. This updated understanding of the relevant features of the world reveals that the methods available to human researchers, GWAS and Twin Studies, are unsuitable for addressing some research questions for which they are currently used, such as questions related to direct genetic effects. If researchers want to overcome current limitations in their methods (i.e., methodological shortfall) to address research questions related to mechanisms, biological pathways, and predictions, then they need to be willing to draw on insights from research areas where sufficient methods are available, like model organism research. This could take the form of increased collaboration between human and model organism researchers. However, this gives rise to challenges because examining the model organism literature may undermine existing commitments, such as the equal environment assumption and little epistasis. Yet this is not a necessary feature of adopting this strategy; rather, it is an empirical question. It may be the case that for some traits these features do not matter. But research in humans will be strengthened by interrogating if this is the case rather than assuming it is not.

3.5 Broader Implications & Conclusions

Having seen the concept of methodological shortfall illustrated in human BG, we can briefly reflect on what this concept can contribute to broader discussions within the philosophy of science. First, methodological shortfall is compatible with other explanations of counterintuitive patterns of activity in science (Section 3.2a), such as the persistence of scientific theories after being faced with negative evidence and the ways that the structure of research questions can blind researchers to possible answers. Instead, methodological shortfall opens a new avenue for investigating how the suitability or sufficiency of available methods shapes the behavior of sciences. While outside the scope of this chapter, there are likely to be interesting connections between methodological shortfall and the breakdown of paradigms and how research questions shape inquiry.

Although I have shown methodological shortfall to be a productive concept for making sense of counterintuitive behavior in human BG, it is a concept that requires intensive knowledge of a specific discipline for it to be recognized; detecting methodological shortfall requires substantial effort. It requires a clear sense of the reach of available methods and an understanding of how these methods are used by the discipline. Both are necessary before one can be positioned to recognize methodological shortfall. Thus, methodological shortfall is not something that can be easily detected by glancing at a discipline's current literature.

Despite the effort required to detect methodological shortfall in an area of science, it can be a useful tool in the arsenal of philosophy of science. Consider contemporary life

sciences. Associated disciplines have gained access to increasing amounts of data that can be acquired from certain automated or high-throughput methods. This has led some researchers to think that access to more data alone will answer outstanding research questions in life science. However, methodological shortfall in BG shows that merely accessing more data from the same methods may still not close the gap in our understanding. In fact, our discussion of BG shows that overinvestment in gathering more data (such as conducting larger GWAS studies) will not only be inadequate to overcome some instances of methodological shortfall, it might also mislead scientists into thinking they are making empirical progress on their research questions.

Within this chapter I have distinguished methodological shortfall from existing explanations of sciences operating counterintuitively. Contemporary BG served as a case study to demonstrate how the concept of methodological shortfall can help us make sense of ongoing issues with contemporary BG research performed in humans. Analyzing these issues through the lens of methodological shortfall revealed two possible paths forward for human BG research: (1) tinker with observational studies like GWAS and Twin Studies to improve them, or (2) adopt or leverage experimental methodologies that can only be used in model organisms, which means approaching their object of study (i.e., behavioral differences) in an indirect way. The latter is the only way to genuinely overcome methodological shortfall in this area of science. Experimental BG with model organisms has the potential to generate epistemic progress if adopted or incorporated by those working in human populations. Results from these experimental quantitative genetic methods can be used to update human BG researchers' understanding of the relevant features of the

world that their methods need to track to provide discriminating tests between competing hypotheses that address their research questions. Current methods alone are not merely insufficient but unsuitable to answer these questions about the causal basis of differences in behavior.

Chapter 4: Does Behavioral Genetics Offer Solutions to Our Social Justice Problems?

4.1 Introduction

"Genetics matter for understanding social inequality" (Harden 2021a, 25). This is the message that readers are left with after finishing *The Genetic Lottery*, a work of public science communication written by behavioral geneticist Kathryn Paige Harden. This book comes at a time when direct-to-consumer genetic testing is increasingly accessible and increasing social inequality is palpable. Harden argues that social science researchers from fields such as anthropology, economics, political science, and sociology are currently blind to an important driver of different forms of social inequality like socioeconomic status, mental health, and educational attainment: genetics. This is a common refrain among behavioral geneticists. Researchers argue that behavioral genetics (BG) can offer social science researchers something other methods cannot – namely, insight into genetic influences on social inequality. For example, behavioral geneticists working on the genetics of socioeconomic status argue that,

> most studies within the broader social sciences aiming to understand social inequalities tend to focus on societal factors. By not including genetic effects, a significantly contributing force is omitted that may be increasing in importance due to recent societal changes (Abdellaoui et al. 2025, 2).

Even general articles meant to explain the uses and limitations of behavioral genetic approaches advertise their utility for investigating social issues like wellbeing. "[T]he

capacity to more clearly illuminate causal pathways for health and wellness outcomes ranks high among what behavioral genetics has to offer the scientific enterprise" (Willoughby, Polderman, and Boutwell 2023, 14).

One proposed use of BG relies on techniques that provide a predictive measure of how likely people are to suffer from poor outcomes such as low socioeconomic status, low educational attainment, and poor mental health based on their genetics. Having this information could allow researchers to identify people who are most likely to perform poorly and offer them additional support. The hope and promise are that incorporating genetic information would increase researchers' ability to identify those who may be struggling, better connect them with the type of help that would most benefit them, thereby reducing social inequality. Given that behavioral geneticists advertise their work as a way to address social inequality, it is worth engaging with the question underlying its basic premise: Should we use behavioral genetics to address social inequality?

This is a *moral question*. It is about whether a course of action should be pursued to provide benefits and reduce harms for individuals. Importantly though, this moral question presumes background knowledge related to the epistemic features of BG methods and results. For any intervention that might be pursued, one must weigh the risks of implementing the intervention with the proposed benefits it may bring. To do this requires knowledge of the pertinent methods and the results purportedly derived from them. Among other potential uses, some behavioral geneticists believe that, if they can use claims about genetics derived from their methods to predict how a person will fare on measures of social inequality, then researchers can devise more effective ways to prevent or remedy

inequality. In these discussions, social inequality encompasses a diversity of phenomena, from economic inequality, disparities in health care, access to mental health resources, and educational achievement. Critically salient features of this question rest on relevant background knowledge. This relevant background knowledge includes but is not limited to: (1) what counts as a behavioral trait for the purpose of scientific research? (2) What are the available BG techniques? (3) What kinds of results do existing methods generate? (4) What are the known limitations of these techniques? These questions fall within the domain of philosophy of science and were the subject of the first three chapters of my dissertation.

I cannot address the potential costs and benefits of using BG to address all forms of social inequality. The strongest arguments from behavioral geneticists in favor of using genetics to address social inequality have focused on the context of education. I will therefore concentrate my argument on the epistemic merits of the proposed use of BG in this context. I argue that BG brings little benefit (and potentially some harm) to the fight against social inequality. In this respect, I am aligned with other bioethicists who have argued for caution and patience regarding the use of BG techniques in education (Matthews 2024; Asbury 2015; "Ethics, Genomics and Education: Findings from a Workshop Exploring the Ethical Landscape of Using Polygenic Indices in Education," n.d.). However, I differ from other bioethicists in my assessment of what BG research would need to look like before it could offer meaningful traction on social justice issues like educational attainment. I leverage analyses from my previous chapters to identify two necessary conditions that BG would need to meet before the bioethics community could re-evaluate whether it should be used to address social inequality. I then consider if a future

field of BG, one that has implemented these conditions, would have anything to offer. I leave open the possibility that the field may one day have something to contribute toward efforts to mitigate social inequality, but I highlight the potential harms of discrimination that are currently left unaddressed, and which may be exacerbated if BG is utilized in the interim.

4.2 Where are we now?

4.2.1 Genetics and Education

Using genetics to understand and potentially improve educational outcomes is not a new idea. Francis Galton pioneered the study of behavioral similarities and differences between family members, and how behavioral performance showed inheritance patterns in families (Galton 1865; Galton 1891). Genetics has come a long way since then. Researchers now have access to genetic sequences, the series of DNA base pairs that make up the heritable material passed down through generations. Researchers also have developed specific methods, such as quantitative genetics, to address specific research questions related to evaluating the genetic contribution to behavioral differences in traits such as educational attainment. While the focus of this chapter is on BG (and its corresponding techniques), it is important to note that there is a history of advocating for the use of a variety of classical and molecular genetic techniques to improve educational outcomes for children in special education programs (Hodapp and Fidler 1999).[17]

[17] Here I focus on cases where researchers have argued that genetic research or genetically informed research could offer benefits to existing children. I do not discuss the lengthy history of eugenic proposals

The causes of children's developmental or intellectual disabilities are heterogenous. It is not unreasonable to think that understanding the cause of a student's intellectual disability could help teachers, parents, and others to potentially improve educational outcomes. For a subset of special education students, genetics has been demonstrated to play an outsized role in their struggles. For example, there are monogenetic (single gene genetic disorders) and chromosomal conditions (e.g., deletions or duplications) that impact an individual's educational attainment by causing developmental delays and impairing intellectual capacities. Advances in molecular genetics and karyotyping have allowed clinicians to identify different ways that intellectual disabilities can arise due to genetic differences. For example, Down Syndrome, which is frequently accompanied by developmental and intellectual delays, is caused by having an additional copy of chromosome 21 ("Down Syndrome - Symptoms and Causes," n.d.). In contrast, phenylketonuria (PKU), another condition that is often accompanied by intellectual and developmental disabilities, is an autosomal recessive monogenetic disorder ("Phenylketonuria (PKU) - Symptoms and Causes," n.d.). PKU is caused by the inability to enzymatically process and remove phenylalanine from the body due to the inheritance of two defective copies of the PAH gene. Down Syndrome and PKU have different underlying genetic etiologies, both of which have been made accessible through classical and molecular genetic techniques. The differences in the etiology of these conditions shape the ways one can intervene to help those who have the conditions.

for the removal of "genetically inferior" individuals from the population to improve population outcomes (often described in terms of national or racial outcomes).

In the case of PKU, molecular genetic sequencing has helped researchers to understand how different mutations in the PAH gene affect the ability to break down phenylalanine. Although PKU is known to be inherited in an autosomal recessive fashion, testing for PKU is done via a blood test of phenylalanine levels. While the incidence rate of PKU in the United States is low, about 1 in 10,000 to 15,000, every newborn in the United States is screened at birth via a blood test to detect levels of phenylalanine in the blood (Strisciuglio and Concolino 2014). This screening enables early intervention to prevent intellectual and developmental disabilities. Once children who may develop PKU have been identified, researchers can implement proven interventions, including special diets low in phenylalanine, additional vitamins, and pharmaceuticals that help to break down phenylalanine.

Although PKU is a genetic disorder that causes intellectual and developmental disability (contributing to educational inequality), physicians do not use genetic techniques to mediate this educational inequality. Instead, they use blood tests and environmental interventions, in the form of diet modifications, to prevent intellectual disabilities. Clearly, understanding the causes of intellectual disabilities can help their mitigate or eliminate their effects. However, within many special education settings, educators have argued that it is not the cause of the child's struggle that matters. Rather, the focus of educational programs generally varies according to the degree of the student's developmental delay or disability, not their etiology. "[T]he cause of mental retardation has little relevance in planning an educational program" (Blackhurst and Berdine 1993, 425). What matters is how students are currently struggling. This indifference to causes of special needs in education is

reflected in the fact that genetics has been rarely used to inform interventions in special education programs (Hodapp and Fidler 1999).

Despite this history, those involved in education, including parents, educators, and education researchers, are now facing renewed pressure to incorporate genetics, in the form of BG, to improve educational outcomes. This pressure comes in the form of notable works of public science like *The Genetic Lottery* and direct-to-consumer genetic testing available through companies like Orchid ("Ethics, Genomics and Education: Findings from a Workshop Exploring the Ethical Landscape of Using Polygenic Indices in Education," n.d.; "Orchid |Whole Genome Embryo Report," n.d.). There are at least three ways these recent calls for using genetics in education differ from previous discussions: a focus on the etiology of behaviors of interest, a difference in the genetic techniques used, and the nature of the population in view.

Arguments for the use of BG in education are not focused on how to tailor interventions to benefit children with monogenetic genetic disorders or chromosomal abnormalities. Rather, individuals like Harden argue for the use of BG to help uncover why individuals differ in *educational abilities in general* (as measured by behavioral performance on evaluative instruments like IQ tests). This use of genetics in education has been referred to as "precision education" or "personalized education," taking inspiration from similar movements in medicine such as "precision medicine" (Matthews 2024; 2020; Tabery 2023). It represents a shift in thinking about educational intervention that comes from a shift in understanding how genetics affect behavior.

BG researchers are interested in identifying how much of the variation in cognitive ability (again, as measured by behavioral performance on evaluative instruments like IQ tests) in a population is due to genetic variation among individuals in that population. By understanding how genetic variation contributes to variation in abilities that contribute to educational outcomes, researchers hope to use this knowledge to improve situations of educational inequality. The problem is that most traits related to educational attainment or outcome like reading ability, math ability, ADHD, dyslexia, personality, and mental health are so-called "polygenic traits." That is, these traits are the results of the combined effect of many hundreds (if not thousands) of genes, each contributing very little to the trait in question. This means that the relationship between genetics and polygenic traits pertinent to educational attainment is far more complicated than the relationship between genetics and a monogenetic condition such as PKU.[18]

There are at least three distinct ways that BG could be used to address educational inequality. The first is to use results of BG studies in the form of a PS[19] to identify students most at risk for poor educational outcomes or those most likely to benefit from a proposed mitigation strategy (Burt 2024; Harden 2021a). This technique would be implemented by collecting students' genetic information and assessing how well, based on their genetic profile (i.e. PS), they would be likely to perform in educational settings. Children predicted to be in the bottom 10-15% of their class could be identified to receive additional

[18] Notice that the scope of the use of genetics in education has also broadened considerably. Before, genetics was being used to target individuals in special education populations who had rare genetic disorders. Now, the new target for precision education is the general student.

[19] As described in Chapter 3, a polygenic score (PS) provides a number that is an aggregated measure of the genetic propensity for some behavior of interest based on thousands of genetic variants of small effect. This measure is sometimes referred to as a Polygenic Risk Score (PRS) or Polygenic Index (PGI).

supportive measures, such as one-on-one tutoring, additional instruction time, or other types of educational assistance. This is similar to traditional risk assessments used in education where measures of reading and math ability are administered throughout the year and children who rank below 10-20% on these measures are identified for additional help. On the face of it, this may not seem that different from the use of other risk assessment measures, like adverse childhood event measures that use a child's family history of major life stressors (e.g., documented physical abuse) to identify children in need of extra educational support. The difference with PS is that children would not be identified based on their current performance or family history—what they can do or what happened to them—but rather by their predicted potential (how they might perform) based only on DNA, which does not vary as the child develops even though behavioral capacities related to educational performance do vary and change throughout development.

The main benefit of this approach would be that educators could identify at-risk students *before* they start to struggle. Traditional measures identify children likely to struggle only after they begin to struggle. Thus, PS could allow educators to intervene earlier, and early interventions have been shown to improve educational outcomes (Shero et al. 2021). Importantly, researchers can develop PS for any behavior of interest, including reading ability or math ability, as well as complex combinations of behavioral performance over time, like general educational attainment. Another possible benefit of this kind of assessment is that it would remove the responsibility of assessment from educators to those processing genetic samples, freeing time and resources for educators already burdened with many other responsibilities. Additionally, genetic testing is often much cheaper than

traditional forms of assessment, which can be time intensive and financially unfeasible to implement at larger scales (Shero et al. 2021).

The second proposal would be to use BG techniques to identify genetic causes of variation in education-related behaviors to better understand biological pathways or mechanisms (Burt 2024; Harden 2021a; Willoughby, Polderman, and Boutwell 2023). For many behaviors that affect educational attainment, researchers do not understand how they arise or how best to intervene. One idea is that BG could be used to identify biological pathways or mechanisms that are relevant to educational outcomes and thereby better understand why things go wrong. In the case of PKU, knowing what biological pathway is affected has allowed researchers to understand that the intellectual and developmental disabilities were caused by a buildup of phenylalanine. Intervening on this pathway, either by removing phenylalanine from the diet so it would not accumulate in the body or artificially performing the function of the PAH gene and removing phenylalanine from the body, prevents affected children from developing an intellectual or developmental disability. BG could help researchers identify the genetic differences that contribute to differences in educational achievement.

The third proposed use of BG is to identify effective interventions in specific research studies. For example, social scientists may want to increase reading proficiency but may be unsure what kind of intervention is most effective. When researchers conduct studies of efficacy, the results are often noisy. Some children seem to respond to the intervention, whereas others do not; and even those that do respond may differ in how much they improve. Adding BG information to social science studies may allow

researchers to better understand these heterogeneous results. It may be that the noise is due to differences in the genetic background of the children. Children with a certain genetic background might respond well to an intervention, while children with other genetic backgrounds might not. With this additional knowledge, researchers might identify that the intervention only works for people with a specific genetic background. This strategy is one of the motivations for the New Haven Lexinome Project, a study run by Yale University researchers. This project aims to "precisely match" those who have dyslexia "with their optimal intervention" ("A More Precise Approach to Dyslexia," n.d.). Yale researchers have collected DNA from students identified to be below the 20th percentile in reading. These students then get access to an hour of intensive reading instruction per day for almost four years. Study leaders argue that children with certain genetic backgrounds may respond better to interventions than others. Results from these studies are not yet available.

All these proposed uses hinge on the epistemic merits of BG methods used in humans. Some researchers, like Harden, argue that BG methods can now be used in the ways suggested above, especially as controls for genetic variation (Harden and Koellinger 2020). Researchers can construct a PS for any education-related behavior for which there is a large enough GWAS. However, the fact that researchers can construct these measures does not mean they are sufficiently accurate or demonstrably useful in educational settings. As will be discussed below, there is significant work to be done before BG has anything useful to offer those interested in mitigating educational inequality.

4.2.2 Limitations

Bioethicists and researchers are keenly aware of some of the methodological limitations of BG (Matthews 2024; Asbury 2015; "Ethics, Genomics and Education: Findings from a Workshop Exploring the Ethical Landscape of Using Polygenic Indices in Education," n.d.). The consensus among those working on the ethics of precision education is that science cannot yet deliver the kind of information needed to implement any of the proposed options discussed above. I provided an extensive analysis of the limitations of BG methods deployed in humans in chapter 3. Here I will briefly summarize the methodological limitations that have featured most prominently in bioethical discussion of precision education.

One commonly cited limitation of current behavioral genetics methods is the issue of interpretation and utility of PSs. Many proponents for the use of PSs interpret them as an *individual* genetic propensity for a given trait. This way of conceptualizing PSs frames them as a measure of the genetic potential of a specific individual. However, a number of researchers argue this is not the correct way to view PSs. (The details of why this is the case can be found in my discussion of methodological limitations in Chapter 3). The harms of using current PSs as measures for individual genetic potential have been summarized by the bioethicist and philosopher of science Lucas Matthews. For example, PSs are very poor predictors of math performance. They are such a poor predictor that, "it is inevitable that many children born with below-average PGS [PSs] for mathematics will grow up to

perform normally or better in mathematics and, vice versa, many children born with above-average scores will grow up to perform normally or worse" (Matthews 2024, 4).

An additional interpretation issue for PSs lies in viewing them as measures of *genetic* propensity or *additive genetic* effects. From our discussion in Chapter 3, it should be clear that the methods available to human researchers cannot distinguish between additive genetic effects and other confounding factors. Therefore, PSs are not a measure of genetics alone, but rather a measure of the mixture of genetic and environmental influences. For this reason, they cannot be used as a measure of direct genetic effects. This rules out the possibility of using PSs to control genetic variation in research studies. Furthermore, PSs are a measure of both genetic and environmental influences. As a consequence, contemporary BG techniques would, "identify schools with a much higher proportion of lower income students from less educated families as having lower *genetic* potential" (Burt 2024, 13, emphasis in original). Instead of attributing students' lower performance on educational tasks to systemic inequality of educational resources, students could instead be labeled as having low genetic potential. This, in turn, could result in schools with more of these students not being identified as 'underperforming' because their students just 'lost' in the 'genetic lottery'"(Burt 2024, 13). In short, students would not be seen as underperforming due to systemic inequality but rather performing as expected given their genetic potential. This label could further *limit* resources available to mitigate inequality, leading to harms rather than benefits.

Even if PSs were a measure of direct genetic effects alone, critics have argued that they should not be understood as a genetic propensity because the direction and strength of the effect can vary due to environment factors (Burt 2024). Recall our discussion of norms of reaction studies in non-human model organisms from Chapter 1. These studies showed that when individuals with the same genetic background were placed in different environments, they exhibited a variety of phenotypes. Although these studies are impossible to perform in humans, the possibility that this pattern applies means that it is possible a child with a low genetic propensity from a PS may do poorly in one environment but actually may do very well in a different environment. Thus, a PS should neither be interpreted as a measure of an *individual* or of a distinctly *genetic* propensity.

Current PS measures have demonstrated little utility. For example, they are currently useless for predicting individual ability in reading and math (Matthews 2024). What miniscule predictive ability PSs have for educational attainment adds remarkably little when used in conjunction with existing measures for predicting student performance. Theoretical models based on existing knowledge provided by BG indicate that there is only a marginal (3%) improvement in identifying "at risk" individuals using results from BG in combination with other interventions (Shero et al. 2021). These predictive measures are less accurate when students are young, undermining one of the main motivations for this option—identifying students at risk of struggling before they actually struggle.

The most widely discussed issue with using contemporary BG methods to address educational inequality is the issue of portability. Portability is a term used in the BG

literature that refers to the fact that PSs and other results are far less accurate when applied to populations that differ from the study population used to construct the measure (Matthews 2022). This means that measures produced in one population cannot be generalized to other populations. For example, most BG studies are conducted in populations of European ancestry, making PSs entirely useless in non-white populations (Mills and Rahal 2019). However, it is precisely the ability to generalize results from one population to another (i.e., portability) that is needed to justify the use of BG results in educational settings.

Behavioral geneticists could attempt to address this problem in part by diversifying their study populations and including non-white individuals (among others). However, diversifying the study population would do nothing to address racial disparities in educational attainment given the current state of the science. This is because diversifying the study population does not address the underlying methodological issues discussed in Chapter 3, especially the inability to distinguish between different kinds of genetic and environmental influences and their resulting interactions. Mitigating the issue of portability does not solve the problem of PSs not representing an individual's genetic potential nor does it address their poor performance as predictors of behavioral performance in education.

Based on a lack of efficacy alone, contemporary BG should not be used to address educational inequality. This includes the use of direct-to-consumer genetic testing that is available to parents and educators through for-profit companies. These companies claim to

offer PSs for intelligence, schizophrenia, and other behaviors relevant for education (including a compound measure of educational attainment). Rejecting the use of PSs for educational intervention is not a controversial stance. Most bioethical discussions of precision education conclude with calls for geneticists to be "extremely careful and humble in communicating both their strengths and limitations of their results" (Matthews 2024; "Ethics, Genomics and Education: Findings from a Workshop Exploring the Ethical Landscape of Using Polygenic Indices in Education," n.d.). Although contemporary BG cannot be used in this way, one might wonder if an improved version could be. What might a future field of BG look like? I leverage work from my earlier chapters to identify two conditions that BG would need to meet before the bioethics community could reevaluate whether PSs or other results from BG methods should be used to address social inequality. Meeting these conditions could both improve the epistemic merits of the field and lead to increased humility and awareness of the field's ongoing limitations.

4.3 What would need to change?

A major argument for the use of any technology is its effectiveness—that it works well or better than other available options. Bioethicists can only reassess whether BG should be used to address educational inequality if the underlying science improves. As described in Chapter 3, behavioral geneticists have tinkered with their existing methods in an attempt to overcome current limitations. This tinkering includes conducting studies with larger sample sizes and diversifying sample populations (e.g., conducting research in populations with non-European ancestry). These forms of tinkering have not overcome the

methodological limitations in human BG, and its track record over time strongly suggests methodological shortfall. For the purpose of this chapter, this means that these changes will not improve the underlying science of BG sufficiently for it to provide much benefit in mitigating social inequality, such as with respect to educational attainment. However, in the previous two chapters, I identified several further areas for methodological improvement. In the following two sections, I demonstrate how these yet-to-be pursued methodological improvements bear on the potential use of BG to mitigate educational inequality.

4.3.1 Traits & Education

Chapter 2 was motivated by the observation that critics of Harden, like Marcus Feldman and Jessica Riskin, objected to the use of these genetic methods to study phenomena like educational attainment, social mobility, and personality that are relevant to social sciences, but not to their use in studying phenomena such as macular degeneration. I argued that the disagreement revolved around different views of what counts as a behavioral *trait* for the purpose of scientific investigation. In the eyes of Feldman and Riskin, some of the behavioral phenomena social scientists are interested in studying are not understood well enough to be considered traits in GWAS (and other studies). Some researchers who share this negative view employ a mechanistic strategy for conceptualizing traits, especially in studying model organisms. That is, they conceptualize behavioral traits like learning and memory in *Drosophila* in terms of underlying shared molecular mechanisms, as is sometimes done with morphological traits like limbs. However, many BG researchers take social science phenomena like socioeconomic status and educational attainment to be traits

for the purpose of their research. These researchers subscribe to something more like conventionalism, which supposes that behavioral traits (and traits more generally) are simply those that can be measured. For example, since researchers have an accepted standardized method of measuring educational attainment in the form of EduYears, it is a behavioral trait.

As discussed in chapter two, this conventionalist strategy of characterizing traits also extends to many other behaviors that researchers study, because they are related to individuals' ability to succeed in education, such as reading ability and math ability.[20] In these studies, reading and math abilities are assessed using standardized tests which yield measures based on behavioral tasks of reading comprehension, reading proficiency, phonics, and general math ability (Haworth et al. 2007). In studies where researchers have more time and resources, such as Twin and Family Studies, they also incorporate qualitative measures developed from reports of parents and teachers (Selzam et al. 2017). Researchers recognize that many behaviors related to education can travel together. For this reason, many studies also include measures of behaviors that are thought to be linked to their behavior of interest. For example, researchers interested in dyslexia will also

[20] Some readers may object that these phenomena are not what we are referring to when we normally talk about behavior. Instead of referring to these as behaviors, we should think of them as abilities, skills, or capacities. Their view is that in referring to these features as behaviors, I am somehow incorrectly lumping them together with phenomena studied in model organisms. However, I disagree. Many researchers working on the genetics of behavior in model organisms study capacities, like the ability to remember learned associations (a behavior I discussed in detail in chapter 2). Additionally, abilities, skills, or capacities in both humans and non-human animals must be measured via the performance of particular behavioral tasks. Cognitive ability, as captured by an IQ test score, is derived from answering questions on the test (i.e., behavioral performance). By referring to the phenomena studied by BG researchers as education-related behaviors, I am drawing an analogy between behaviors in humans and behaviors in non-human organism that is well-motivated by how these are actually measured.

include standardized tests of behaviors like word reading, non-word reading, spelling, and digit span[21] (Gialluisi et al. 2019).

However, being able to measure behaviors does not mean the behaviors are well conceptualized as traits. Take reading ability as an example. Researchers disagree on what kind of measures constitute poor reading ability. Some argue that poor reading only applies to those who rank below the 10th percentile in a measure of reading ability. Others disagree, saying that poor reading ability should apply to all who rank below the 20th percentile. This debate encompasses more than merely the threshold cut-off. Many of the same measures of reading ability are used to assess dyslexia in GWAS (Price et al. 2022; Luciano et al. 2013; Price et al. 2020; Eising et al. 2022; Gialluisi et al. 2014; Gialluisi et al. 2019). We may want genetic studies of reading ability to help us distinguish between poor reading due to dyslexia and poor reading due to other sources. Even within genetic studies of dyslexia, the level of conceptualization for the relevant behavior varies widely. An example of poor conceptualization for dyslexia as a trait can be found in a GWAS utilizing the 23 and Me database. This study merely used a positive answer to the question "have you ever been diagnosed with dyslexia?" as its inclusion criterion for a GWAS of dyslexia (Doust et al. 2022). Although researchers do employ measures of behaviors in their studies, close examination of how these measures are used reveals key pitfalls of using only a conventionalist understanding of behavioral traits in BG studies.

[21] A test of one's ability to remember sequences of numbers after presentation.

As I argued in Chapter 2, conventionalism is too permissive a standard to be used in the genetic study of behavior because it licenses almost anything as a trait, including artifacts or gerrymandered categories. However, the other existing alternative (the mechanistic strategy) is too stringent, virtually ruling out most behaviors as traits, especially in human BG investigations. The mechanistic account requires researchers to have existing knowledge of the causal mechanisms for a behavior, a feature currently missing for many of the phenomena of interest to behavioral geneticists. For BG studies of education-related behaviors to be meaningful, more work needs to be done characterizing what facet of the behavior is under study within BG. Since existing strategies are ill-suited for characterizing behavioral traits in genetic studies, I proposed a novel alternative strategy (in chapter 3): the Analogical Contextual Strategy (ACS). The ACS strikes a balance between these extremes, providing a basis for researchers across many different areas of behavioral research to conceptualize traits for study in both human-focused inquiry and model organism research. Because the ACS consists of a suite of analogies from the study of morphology that are neither necessary nor sufficient, it facilitates a shift away from the conceptualization of behavioral traits *in general* to specific traits of interest according to specified criteria.

The ACS can be used to improve the characterization of behaviors of interest for BG, like educational attainment, reading ability, and math ability. I argued in Chapter 2 that educational attainment should not be considered a behavioral trait for the purpose of scientific investigation into genetic influences on behavior because it is not sufficiently characterized as a trait. Educational attainment may be more accurately described as a

proxy for the phenomenon researchers are interested in, such the way points scored by dunking in basketball could be a measure of leaping ability. Although educational attainment failed to meet all four criteria of the ACS, the most relevant criterion for discussions of educational inequality was its failure to be well characterized in terms of biological function. For most social science research, characterizing behavior in terms of biological function is not an issue because it is not relevant to most of their research questions. For example, a deep notion of function is not needed for social science researchers to use educational attainment to track the impact of economic or social inequality. However, this criterion becomes relevant for questions where researchers have an implicit commitment to mechanisms. This commitment is present in BG and associated intervention recommendations like precision education, because there must be some pathway that connects genetic variation to variation in education-related behaviors.

Recognizing a behavior as poorly characterized with respect to specific parameters does not mean research on these features cannot continue. Rather, it is an invitation to engage in research that facilitates better conceptualization along these lines. One consequence of employing the ACS is that it reveals specific areas where the conceptualization of a behavior can be improved. Researchers can then prioritize research in these areas. For example, BG research on reading behavior often does not distinguish between poor reading ability and dyslexia. While these behaviors can be correlated with one another, there may be many research questions where distinguishing between them is important. Until the relationship between these behaviors is better conceptualized, researchers could focus their research on those features that are shared by both behaviors.

This could include conducting studies of how genetic variation is related to variation in the ability to remember a string of digits and to accurately perform phoneme recognition or spell words. These behaviors are still relevant to the student's ability to succeed in education, but they are better conceptualized using the four considerations from the ACS. This would then make them better candidates for a study using BG methods. Additional research could focus on whether those with dyslexia or poor reading ability due to other causes differ meaningfully regarding these measures.

Researchers can study education-related behaviors like reading ability that contribute to educational attainment. However, this does not mean that these behaviors are well-conceptualized traits. Viewing these behaviors as traits implies that researchers are capturing something meaningful in an investigation and that they have a good idea of what it is they are capturing. This is not the case for reading ability, and especially not the case for proxies like educational attainment. Therefore, researchers need to hold on to them lightly. How scientists and researchers conceptualize behaviors changes over time, and it may be that with future research there will be a need to reevaluate the trait status of many behaviors related to education attainment. The process of investigating education-related behaviors along the lines suggested by the ACS is crucial. Unless deeper conceptual work on behavioral traits is accomplished, BG will have little to offer the investigation of biological pathways and mechanisms that underlie educational inequality.

4.3.2 Methodological Shifts

Bioethicists have called for human BG to be humble and cognizant of its limitations. My discussion in Chapter 3 offered some reasons why hubris and a lack of awareness persist. Human BG exhibits methodological shortfall; its existing research methods are unable to discriminate between competing hypotheses that bear on key research questions, including questions relevant to precision education. I argued that although researchers have tinkered with existing methods, and that this has augmented the ability to discriminate between competing hypotheses to some degree, the core methodological limitations are still present. I also demonstrated that the field has maintained optimism in its ability to soon overcome these limitations by adopting new technologies. This technological optimism is one of the drivers of behavioral geneticists' lack of humility even though observational methods using these new technologies are unable to address problems like the confounding of genetic and environmental factors in heritability estimates for behavioral trait differences.

I also demonstrated in Chapter 3 how experimental quantitative genetic methods used in non-human model organism research can overcome this shortfall and provide answers to central research questions. In doing so, results from these studies provide the opportunity to update the understanding of researchers in human BG with respect to the relationship between genetic variation and environmental variation. Insights from model organisms showed that there are important aspects of the complex interplay between genetics and the environment that cannot be disentangled based on research carried out in humans. These findings are relevant to the use of BG in education because proposed uses

for educational intervention rely on these methods. Research in model organisms shows that current human BG methods are severely limited and these limitations will persist even if researchers diversify study populations and increase the number of individuals included in research studies (i.e., tinkering with their existing methods). This observation suggests a way forward empirically. It also may provide an opportunity for the field to practice humility, a disposition useful for scientific collaboration.

If researchers in human BG want to overcome current limitations in their methods, then they need to be willing to draw insights from research areas where sufficient methods are available, such as model organism research. This could mean increased collaboration between human and model organism researchers focused on BG. Methodological shortfall in human BG shows that merely accessing more data from the same methods will not close the gap in understanding. In fact, our discussion of human BG shows that a continued investment in gathering more data (such as by conducting larger GWASs) will not be enough to overcome key aspects of methodological shortfall. For the field of human BG to have anything useful to offer those interested in improving educational inequality, human and model organism researchers would need to work in concert. This collaboration would bring clear epistemic benefits, including an awareness of how limitations may be improved through strategic use of results from model organism research.

As I described in Chapter 1, BG is an insular discipline. It has closed itself off from other closely related fields, in part to shield itself from criticism. Drawing on the model organism literature will undoubtedly challenge many existing commitments, such as the equal environment assumption and the presumption of low frequencies of epistatic effects.

It may also cause conflict between the different research communities. This may raise distinct sociological challenges related to how these groups interact. Yet the extent to which these interactions challenge existing commitments is not a necessary feature of this collaboration. Rather, it is an empirical question. For some traits, the challenges may be few, and some commitments may survive relatively intact. However, this investigative process may put pressure on the view among some behavioral geneticists that humans are exceptional and that, accordingly, research focused on model organisms is not relevant. In fact, I have made the case that research focused on humans will be strengthened by interrogating whether this is the case rather than assuming that it is not.

4.4 Predicting the future

The field of BG needs to change, and change radically, before it can offer any assistance to addressing educational inequality. The avenues for change that I have suggested make it difficult to predict whether a new and improved field will emerge. I have argued that this notoriously insular and highly defensive field needs to leverage techniques it can only access by working with other research communities. This kind of collaboration will not just require epistemic openness and humility but also revisions to the sociological structure of the discipline. That is, researchers working across taxa will need to learn to see their work as having consequences for those in other taxa. Researchers will need to reorient their investigations away from assuming that humans are different from other organisms and, rather, towards establishing that humans constitute an exception (if this is indeed the case).

How would a more substantial relationship between human and model organism researchers change the current understanding of the field? It may be the case that future research will improve PSs to be more accurate as predictive measures for individuals or at least reduce the risk of inaccurately identifying those most in need of additional educational assistance. Future research may also incorporate diverse study populations and thus allow for BG methods and results to apply to populations of non-European ancestry. Even with these particular risks mitigated, however, some serious risks remain.

4.4.1 What (potential) harm remains?

The strategies described above cannot prevent the risk of some of the harms that may result from implementing precision education (i.e., using BG results, like PSs, in educational interventions). One of these harms is the genetic Pygmalion effect (Matthews 2024). This effect gets its name from work describing how teachers' expectations of students affect students' performance. In short, the Pygmalion effect describes a self-fulfilling prophecy where the expectations one has for a student shape how the student ends up doing in the classroom. There is good evidence for this effect. While the traditional Pygmalion effect is based on an educator's expectation of their students, there is a concern that access to PSs for educators could cause a similar effect – the polygenic Pygmalion effect. The idea would be that PSs will shape the student trajectory in a way that fulfills the prediction of the PS. Teachers will lower their expectations for students who receive PSs indicative of low genetic potential and, conversely, raise their expectations for students who receive PSs of high genetic potential. The results would be detrimental for the former set of students and, potentially, beneficial for the latter, thereby exacerbating educational inequality rather than

reducing it. There is ongoing research in this area. Preliminary studies have shown that people who receive low PS results report a higher incidence of negative feelings of self-esteem and academic achievement (Matthews et al. 2021).

Growing concern surrounding the implementation of precision education has motivated research into developing mitigation techniques to prevent or lessen the incidence of the genetic Pygmalion effect. Although there are very few published studies to date, and the studies that do exist are underpowered, researchers have found that people who are presented with a description of the influence of PSs report higher rates of competence and academic efficacy then those who were given similar information about the educational environment (Matthews, Zhang, and Martschenko 2024). Although the risk of the genetic Pygmalion effect remains, there are encouraging findings that genetic reports paired with additional educational information may be able to mitigate this effect. However, given our current understanding, the risks associated with the polygenic Pygmalion effect remain uncertain.

Another risk that has been consistently described in the bioethics literature related to precision education is the collapsing of a person's potential and value to their genetic profile. I describe this as "genetic synecdoche" because a person's genetics (including PSs) stands in for the whole of who they are (Hall 2013; Parens and Ash 2000; Asch and Wasserman 2005). The risk of synecdoche is well described in the literature for bioethics and critical disability studies where there has been growing concern that the results of genetic testing play an outsized role in parents' choice of embryos or in choosing to continue a pregnancy. The argument is that these kinds of measures reduce the embryo to

a particular diagnosis or set of features rather than taking it as a complex, developing whole. The Nuffield council on bioethics recent workshop on Ethics, Genomics, and Education noted that because genetics is considered immutable, there was concern from participants that this could "contribute to genetic determinism, that is, the erroneous perception that genetic make-up is the sole, or most important, driver of individual outcomes" (14). The risk of genetic synecdoche is that it could lead to the belief that the value and potential of a person are directly tied to their genetics, and in a way immune from any environmental influence. If anything can be taken from model organism research, it should be clear that while genetic sequences are difficult to change, they do not convey fixed potential for most features, especially complex traits like educational attainment. Thus, implementing precision education carries the risk of enforcing mistaken notions that genetics *determines* one's fate or life outcomes. This is simply not the case.

Moreover, implementing precision education carries the risk of perpetuating inequality based on genetic synecdoche. Precision education would operate by accurately identifying students according to their genetic profile (maybe in concert with other measures). In a future where BG could deliver this kind of information (which it currently cannot), science tells us nothing about how it should be used. Proponents of precision education, like Harden, argue that the use of this information can be consistent with egalitarian ethics and is necessary to any attempt to improve people's lives.

The widespread tendency to ignore the existence of genetic differences between people has hobbled scientific progress in psychology, education, and other branches of the social sciences. As a result, we have been much less successful at understanding human

development and at intervening to improve human lives than we could be… If social scientists are collectively going to rise to the challenge of improving people's lives, we cannot afford to ignore a fundamental fact about human nature: that people are not born the same (Harden 2021a, 21).

However, recognizing that people are not born the same, and using their dissimilarity as the basis for how we treat them, always raises the risk that we will discriminate against those whom we see as different. This is especially a risk to those we see as "less than" according to some standard or measure. The same technique that could allow us to identify the students most likely to struggle in school and motivate us to provide access to additional educational support, is the same technique that might encourage us to bar these same students from higher education or provide genetically high-achieving students with even more support. This technique has as much potential to magnify social inequality as it does to mitigate it.

4.5 Conclusion

Within this chapter I have provided a negative answer to the question of whether BG should be used to address social inequality. I first focused my argument on the epistemic merits of the proposed use of BG. I argued that the field currently brings little benefit to the fight against social inequality. In this stance, I am aligned with other bioethicists who have argued for caution and patience regarding the use of BG techniques in education interventions. Where I diverge from other bioethicists is in my articulation of what BG research would need to look like before the field could offer meaningful help with social justice issues like educational inequality. I outlined conditions that BG would need to meet

before one could reevaluate whether it should be used to address one or more forms of social inequality. The changes I suggested the field needs to make would result in a radically different area of science, including revisions to core theoretical commitments like the equal environment assumption or the presumption of a low frequency of epistatic effects. As a consequence, it is difficult to predict whether or when this might happen and how the findings of BG might look afterward. I leave open the possibility that the field may one day have something to offer to those working to mitigate social inequality, but for now the risk of serious harms remains.

Bibliography

"A More Precise Approach to Dyslexia." n.d. Accessed May 13, 2025. https://medicine.yale.edu/news/medicineatyale/article/a-more-precise-approach-to-dyslexia/.

Abdellaoui, Abdel, Hilary C. Martin, Martin Kolk, Adam Rutherford, Michael Muthukrishna, Felix C. Tropf, Melinda C. Mills, Brendan Zietsch, Karin J.H. Verweij, and Peter M. Visscher. 2025. "Socio-Economic Status Is a Social Construct with Heritable Components and Genetic Consequences." | *Cambridge Quarterly of Healthcare Ethics*, March, 1–17.

Abdellaoui, Abdel, and Karin J. H. Verweij. 2021. "Dissecting Polygenic Signals from Genome-Wide Association Studies on Human Behaviour." *Nature Human Behaviour* 5 (6): 686–94.

Alberts, Bruce, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. 2002. "Looking at the Structure of Cells in the Microscope." In *Molecular Biology of the Cell. 4th Edition*. Garland Science. https://www.ncbi.nlm.nih.gov/books/NBK26880/.

Anholt, Robert R. H. 2010. "Making Scents of Behavioural Genetics: Lessons from *Drosophila*." *Genetics Research* 92 (5–6): 349–59.

Anholt, Robert R.H., and Trudy F.C. Mackay. 2004. "Quantitative Genetic Analyses of Complex Behaviours in Drosophila." *Nature Reviews Genetics* 5 (11): 838–49.

Ankeny, Rachel A. 2010. "Historiographic Reflections on Model Organisms: Or How the Mureaucracy May Be Limiting Our Understanding of Contemporary Genetics and Genomics." *History and Philosophy of the Life Sciences* 32 (1): 91–104.

Ankeny, Rachel A., and Sabina Leonelli. 2011. "What's so Special about Model Organisms?" *Studies in History and Philosophy of Science Part A*, Model-Based Representation in Scientific Practice, 42 (2): 313–23.

———. 2021. *Model Organisms*. Elements in the Philosophy of Biology. Cambridge: Cambridge University Press.

Ankeny, Rachel A., Sabina Leonelli, Nicole C. Nelson, and Edmund Ramsden. 2014. "Making Organisms Model Human Behavior: Situated Models in North-American Alcohol Research, since 1950." *Science in Context* 27 (3): 485–509.

Anney, Richard J. L., Jessica Lasky-Su, Colm O'Dúshláine, Elaine Kenny, Benjamin M. Neale, Aisling Mulligan, Barbara Franke, et al. 2008. "Conduct Disorder and ADHD: Evaluation of Conduct Problems as a Categorical and Quantitative Trait in the International Multicentre ADHD Genetics Study." *American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics* 147B (8): 1369–78.

Ariew, André, Robert Cummins, and Mark Perlman. 2002. *Functions: New Essays in the Philosophy of Psychology and Biology*. Oxford University Press.

Asbury, Kathryn. 2015. "Can Genetics Research Benefit Educational Interventions for All?" *Hastings Center Report* 45 (S1): S39–42.

Asch, Adrienne, and David Wasserman. 2005. "Where Is the Sin in Synecdoche? Prenatal Testing and the Parent-Child Relationship." In *Quality of Life and Human Difference: Genetic Testing, Health Care, and Disability*, edited by David Wasserman, Jerome Bickenbach, and Robert Wachbroit, 172–216. Cambridge Studies in Philosophy and Public Policy. Cambridge: Cambridge University Press.

Bechtel, William. 2005. *Discovering Cell Mechanisms: The Creation of Modern Cell Biology*. Cambridge Studies in Philosophy and Biology. Cambridge: Cambridge University Press.

Bergelson, J., Roux, F. 2010. *Towards identifying genes underlying ecologically relevant traits in Arabidopsis thaliana.* Nature Reviews Genetics **11**, 867–879.

Bergner, Raymond M. 2016. "What Is Behavior? And Why Is It Not Reducible to Biological States of Affairs?" *Journal of Theoretical and Philosophical Psychology* 36 (1): 41–55.

Bernardo, Rex. 2008. "Molecular Markers and Selection for Complex Traits in Plants: Learning from the Last 20 Years." *Crop Science* 48 (5): 1649–64.

Berry, Christopher M., Melissa L. Gruys, and Paul R. Sackett. 2006. "Educational Attainment as a Proxy for Cognitive Ability in Selection: Effects on Levels of Cognitive Ability and Adverse Impact." *Journal of Applied Psychology* 91 (3): 696–705.

Bier, Ethan, and William McGinnis. 2008. "Model Organisms in the Study of Development and Disease." In *Inborn Errors Of Development: The Molecular Basis of Clinical Disorders of Morphogenesis*, edited by Charles J Epstein, Robert P Erickson, and Anthony Wynshaw-Boris, 0. Oxford University Press.

Blackhurst, A. Edward, and William H. Berdine. 1993. *An Introduction to Special Education*. New York, NY: HarperCollins College Publishers.

Boggess, Scott. 1998. "Family Structure, Economic Status, and Educational Attainment." *Journal of Population Economics* 11 (2): 205–22.

Bosker, F. J., C. A. Hartman, I. M. Nolte, B. P. Prins, P. Terpstra, D. Posthuma, T. van Veen, et al. 2011. "Poor Replication of Candidate Genes for Major Depressive Disorder Using Genome-Wide Association Data." *Molecular Psychiatry* 16 (5): 516–32.

Boyd, Nora Mills, and Dana Matthiessen. 2024. "Observations, Experiments, and Arguments for Epistemic Superiority in Scientific Methodology." *Philosophy of Science* 91 (1): 111–31.

Bralten, Janita, Nina R. Mota, Cornelius J. H. M. Klemann, Ward De Witte, Emma Laing, David A. Collier, Hilde de Kluiver, et al. 2021. “Genetic Underpinnings of Sociability in the General Population.” *Neuropsychopharmacology* 46 (9): 1627–34.

Branigan, Amelia R., Kenneth J. McCallum, and Jeremy Freese. 2013. “Variation in the Heritability of Educational Attainment: An International Meta-Analysis.” *Social Forces* 92 (1): 109–40.

Brevik, Erlend J., Marjolein M. J. van Donkelaar, Heike Weber, Cristina Sánchez-Mora, Christian Jacob, Olga Rivero, Sarah Kittel-Schneider, et al. 2016. “Genome-Wide Analyses of Aggressiveness in Attention-Deficit Hyperactivity Disorder.” *American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics* 171 (5): 733–47.

Brown, Rachael L. 2014. “Identifying Behavioral Novelty.” *Biological Theory* 9 (2): 135–48.

Brumpton, Ben, Eleanor Sanderson, Karl Heilbron, Fernando Pires Hartwig, Sean Harrison, Gunnhild Åberge Vie, Yoonsu Cho, et al. 2020. “Avoiding Dynastic, Assortative Mating, and Population Stratification Biases in Mendelian Randomization through within-Family Analyses.” *Nature Communications* 11 (1): 3519.

Bulik-Sullivan, Brendan K., Po-Ru Loh, Hilary K. Finucane, Stephan Ripke, Jian Yang, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Nick Patterson, Mark J. Daly, Alkes L. Price, and Benjamin M. Neale. 2015. “LD Score Regression Distinguishes Confounding from Polygenicity in Genome-Wide Association Studies.” *Nature Genetics* 47 (3): 291–95.

Burt, Callie H. 2024. “Polygenic Indices (a.k.a. Polygenic Scores) in Social Science: A Guide for Interpretation and Evaluation.” *Sociological Methodology* 54 (2): 300–350.

Burt, S. Alexandra, Kathryn S. Plaisance, and David Z. Hambrick. 2019. “Understanding ‘What Could Be’: A Call for ‘Experimental Behavioral Genetics.’” *Behavior Genetics* 49 (2): 235–43.

Busto, Germain U., Isaac Cervantes-Sandoval, and Ronald L. Davis. 2010. “Olfactory Learning in Drosophila.” *Physiology* 25 (6): 338–46.

Caldicott, Lyn, Helen E. Zulch, Thomas W. Pike, and Anna Wilkinson. 2023. “Olfactory Learning and Training Methods.” In *Olfactory Research in Dogs*, edited by Lucia Lazarowski, 177–204. Cham: Springer International Publishing.

Cesarini, David, Christopher T. Dawes, Magnus Johannesson, Paul Lichtenstein, and Björn Wallace. 2009. “Genetic Variation in Preferences for Giving and Risk Taking.” *Quarterly Journal of Economics* 124 (2): 809–42.

Charney, Evan. 2012a. “Behavior Genetics and Postgenomics.” *The Behavioral and Brain Sciences* 35 (5): 331–58.

———. 2012b. “Humans, Fruit Flies, and Automatons.” *Behavioral and Brain Sciences* 35 (5): 381–410.

Chen, Hung-Lun, Chien-Fu F. Chen, and Han-Bin Huang. 2022. “Distinct Age-Specific Effects on Olfactory Associative Learning in C57BL/6 Substrains.” *Frontiers in Behavioral Neuroscience* 16 (February).

Chen, Tzu-Ting, Jaeyoung Kim, Max Lam, Yi-Fang Chuang, Yen-Ling Chiu, Shu-Chin Lin, Sang-Hyuk Jung, et al. 2024. “Shared Genetic Architectures of Educational Attainment in East Asian and European Populations.” *Nature Human Behaviour* 8 (3): 562–75.

Coop, Graham, and Molly Przeworski. 2022a. “Lottery, Luck, or Legacy. A Review of ‘The Genetic Lottery: Why DNA Matters for Social Equality.’” *Evolution* 76 (4): 846–53.

———. 2022b. “Luck, Lottery, or Legacy? The Problem of Confounding. A Reply to Harden.” *Evolution* 76 (10): 2464–68.

Davies, G., R. E. Marioni, D. C. Liewald, W. D. Hill, S. P. Hagenaars, S. E. Harris, S. J. Ritchie, et al. 2016. “Genome-Wide Association Study of Cognitive Functions and Educational Attainment in UK Biobank (N=112 151).” *Molecular Psychiatry* 21 (6): 758–67.

Davis, Ronald L. 2023. “Learning and Memory Using Drosophila Melanogaster: A Focus on Advances Made in the Fifth Decade of Research.” *Genetics* 224 (4): iyad085.

Derringer, Jaime, Robin P. Corley, Brett C. Haberstick, Susan E. Young, Brittany A. Demmitt, Daniel P. Howrigan, Robert M. Kirkpatrick, et al. 2015. “Genome-Wide Association Study of Behavioral Disinhibition in a Selected Adolescent Sample.” *Behavior Genetics* 45 (4): 375–81.

Dick, D. M., F. Aliev, R. F. Krueger, A. Edwards, A. Agrawal, et al. 2011. “Genome-Wide Association Study of Conduct Disorder Symptomatology.” *Molecular Psychiatry* 16 (8): 800–808.

DiFrisco, James, Alan C. Love, and Günter P. Wagner. 2020. “Character Identity Mechanisms: A Conceptual Model for Comparative-Mechanistic Biology.” *Biology & Philosophy* 35 (4): 44.

DiFrisco, James, Günter P. Wagner, and Alan C. Love. 2023. “Reframing Research on Evolutionary Novelty and Co-Option: Character Identity Mechanisms versus Deep Homology.” *Seminars in Cell & Developmental Biology*, 145 (August):3–12.

Dobzhansky, T., and B. Spassky. 1944. “Genetics of Natural Populations. Xi. Manifestation of Genetic Variants in Drosophila pseudoobscura in Different Environments.” *Genetics* 29 (3): 270–90.

Dobzhansky, Theodosius. 1956. “What Is an Adaptive Trait?” *The American Naturalist* 90 (855): 337–47.

Doust, Catherine, Pierre Fontanillas, Else Eising, Scott D. Gordon, Zhengjun Wang, Gökberk Alagöz, Barbara Molz, et al. 2022. "Discovery of 42 Genome-Wide Significant Loci Associated with Dyslexia." *Nature Genetics* 54 (11): 1621–29.

"Down Syndrome - Symptoms and Causes." n.d. Mayo Clinic. Accessed May 13, 2025. https://www.mayoclinic.org/diseases-conditions/down-syndrome/symptoms-causes/syc-20355977.

Downes, Stephen M., and Lucas Matthews. 2020. "Heritability." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Spring 2020. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2020/entries/heredity/.

Dubnau, Josh. 2014. *Behavioral Genetics of the Fly (Drosophila Melanogaster)*. Cambridge, UNITED KINGDOM: Cambridge University Press. http://ebookcentral.proquest.com/lib/umn/detail.action?docID=1658757.

Duijn, Marc van, Fred Keijzer, and Daan Franken. 2006. "Principles of Minimal Cognition: Casting Cognition as Sensorimotor Coordination." *Adaptive Behavior* 14 (2): 157–70.

Eaves, Lindon J., and Peter K. Hatemi. 2011. "Do We Choose Our Spouse Based on Our In-Laws? Resolving the Effects of Family Background and Spousal Choice for Educational Attainment, Religious Practice, and Political Preference." *Social Science Quarterly* 92 (5): 1253–78.

Edwards, Alexis C., and Trudy F. C. Mackay. 2009. "Quantitative Trait Loci for Aggressive Behavior in Drosophila Melanogaster." *Genetics* 182 (3): 889–97.

Eising, Else, Nazanin Mirza-Schreiber, Eveline L. de Zeeuw, Carol A. Wang, Dongnhu T. Truong, Andrea G. Allegrini, Chin Yang Shapland, et al. 2022. "Genome-Wide Analyses of Individual Differences in Quantitatively Assessed Reading- and Language-Related Skills in up to 34,000 People." *Proceedings of the National Academy of Sciences USA* 119 (35): e2202764119.

"Ethics, Genomics and Education: Findings from a Workshop Exploring the Ethical Landscape of Using Polygenic Indices in Education." n.d. Nuffield Council on Bioethics. Accessed May 13, 2025. https://www.nuffieldbioethics.org/publication/ethics-genomics-and-education-findings-from-a-workshop-exploring-the-ethical-landscape-of-using-polygenic-indices-in-education/.

Faber, Till, Jasdan Joerges, and Randolf Menzel. 1999. "Associative Learning Modifies Neural Representations of Odors in the Insect Brain." *Nature Neuroscience* 2 (1): 74–78.

Falconer, D. S. 1996. *Introduction to Quantitative Genetics*. 4th ed.. Burnt Mill, Harlow, Essex, England: Longman.

Ferguson, Christopher J. 2010. "Genetic Contributions to Antisocial Personality and Behavior: A Meta-Analytic Review from an Evolutionary Perspective." *The Journal of Social Psychology* 150 (2): 160–80.

Filmer, Deon, and Lant Pritchett. 1999. “The Effect of Household Wealth on Educational Attainment: Evidence from 35 Countries.” *Population and Development Review* 25 (1): 85–120.

Flynn, James R. 1987. “Massive IQ Gains in 14 Nations: What IQ Tests Really Measure.” *Psychological Bulletin* 101 (2): 171–91.

Friedman, Naomi P., Marie T. Banich, and Matthew C. Keller. 2021. “Twin Studies to GWAS: There and Back Again.” *Trends in Cognitive Sciences* 25 (10): 855–69.

Galton, Francis. 1865. “Hereditary Talent and Character.” *Macmillan’s Magazine* 12 ((157-166)): 318–27.

———1891. *Hereditary Genius*. D. Appleton.

Ganna, Andrea, Karin J. H. Verweij, Michel G. Nivard, Robert Maier, Robbee Wedow, Alexander S. Busch, Abdel Abdellaoui, et al. 2019. “Large-Scale GWAS Reveals Insights into the Genetic Architecture of Same-Sex Sexual Behavior.” *Science* 365 (6456): eaat7693.

Gao, Luyan, Kun Wang, Qing-Bin Ni, Hongguang Fan, Lan Zhao, Lei Huang, Mingfeng Yang, and Huanming Li. 2022. “Educational Attainment and Ischemic Stroke: A Mendelian Randomization Study.” *Frontiers in Genetics* 12 (February).

Gialluisi, A., D. F. Newbury, E. G. Wilcutt, R. K. Olson, J. C. DeFries, W. M. Brandler, B. F. Pennington, et al. 2014. “Genome-Wide Screening for DNA Variants Associated with Reading and Language Traits.” *Genes, Brain and Behavior* 13 (7): 686–701.

Gialluisi, Alessandro, Till F. M. Andlauer, Nazanin Mirza-Schreiber, Kristina Moll, Jessica Becker, Per Hoffmann, Kerstin U. Ludwig, et al. 2019. “Genome-Wide Association Scan Identifies New Variants Associated with a Cognitive Predictor of Dyslexia.” *Translational Psychiatry* 9 (1): 1–15.

Gould, S.J. 1981. *The Mismeasure of Man*. 1st ed. New York: Norton.

Gould S.J., Lewontin R.C. 1979. *The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme*. Proceedings of the Royal Society: Biological Sciences. 205(1161):581-98.

Griffiths, Paul E., and James Tabery. 2008. “Behavioral Genetics and Development: Historical and Conceptual Causes of Controversy.” *New Ideas in Psychology*, Integrating development and evolution in psychology, 26 (3): 332–52.

Gummalla, Maheshwar, Sandrine Galetti, Robert K. Maeda, and François Karch. 2014. “Hox Gene Regulation in the Central Nervous System of Drosophila.” *Frontiers in Cellular Neuroscience* 8 (April).

Hagenbeek, Fiona A., Jana S. Hirzinger, Sophie Breunig, Susanne Bruins, Dmitry V. Kuznetsov, Kirsten Schut, Veronika V. Odintsova, and Dorret I. Boomsma. 2023.

"Maximizing the Value of Twin Studies in Health and Behaviour." *Nature Human Behaviour* 7 (6): 849–60.

Hall, Melinda. 2013. "Reconciling the Disability Critique and Reproductive Liberty: The Case of Negative Genetic Selection." *International Journal of Feminist Approaches to Bioethics Vol* 6 (1): 121–43.

Harden, K. Paige. 2021a. *The Genetic Lottery: Why DNA Matters for Social Equality*. Princeton University Press.

——— 2021b. "'Reports of My Death Were Greatly Exaggerated': Behavior Genetics in the Postgenomic Era." *Annual Review of Psychology* 72 (January):37–60.

——— 2022. "On Genetics and Justice: A Reply to Coop and Przeworski (2022)." *Evolution* 76 (10): 2469–74.

Harden, K. Paige, and Philipp D. Koellinger. 2020. "Using Genetics for Social Science." *Nature Human Behaviour* 4 (6): 567.

Haworth, Claire M. A., Nicole Harlaar, Yulia Kovas, Oliver S. P. Davis, Bonamy R. Oliver, Marianna E. Hayiou-Thomas, Jane Frances, et al. 2007. "Internet Cognitive Testing of Large Samples Needed in Genetic Research." *Twin Research and Human Genetics* 10 (4): 554–63.

Hagenbeek, Fiona A., Jana S. Hirzinger, Sophie Breunig, Susanne Bruins, Dmitry V. Kuznetsov, Kirsten Schut, Veronika V. Odintsova, and Dorret I. Boomsma. 2023. "Maximizing the Value of Twin Studies in Health and Behaviour." *Nature Human Behaviour* 7 (6): 849–60.

Henriques, Gregg, and Joseph Michalski. 2020. "Defining Behavior and Its Relationship to the Science of Psychology." *Integrative Psychological and Behavioral Science* 54 (2): 328–53.

Herrnstein, Richard J. 1994. *The Bell Curve: Intelligence and Class Structure in American Life*. New York: Free Press.

Hill, W. David, Neil M. Davies, Stuart J. Ritchie, Nathan G. Skene, Julien Bryois, Steven Bell, Emanuele Di Angelantonio, et al. 2019. "Genome-Wide Analysis Identifies Molecular Systems and 149 Genetic Loci Associated with Income." *Nature Communications* 10 (1): 5741.

Hodapp, Robert M., and Deborah J. Fidler. 1999. "Special Education and Genetics: Connections for the 21st Century." *The Journal of Special Education* 33 (3): 130–37.

Howe, Laurence J., Michel G. Nivard, Tim T. Morris, Ailin F. Hansen, Humaira Rasheed, Yoonsu Cho, Geetha Chittoor, et al. 2022. "Within-Sibship Genome-Wide Association

Analyses Decrease Bias in Estimates of Direct Genetic Effects." *Nature Genetics* 54 (5): 581–92.

"International Standard Classification of Education (ISCED) - Statistics Explained." n.d. Accessed June 19, 2024. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=International_Standard_Classification_of_Education_(ISCED).

Jaffee, S. R., and T. S. Price. 2007. "Gene–Environment Correlations: A Review of the Evidence and Implications for Prevention of Mental Illness." *Molecular Psychiatry* 12 (5): 432–42.

Jensen, A. 1973. "Educability and Group Differences". [1st U.S. ed.]. *New York: Harper & Row*.

Linnér ,K., Richard, Pietro Biroli, Edward Kong, S. Fleur W. Meddens, Robbee Wedow, Mark Alan Fontana, Maël Lebreton, et al. 2019. "Genome-Wide Association Analyses of Risk Tolerance and Risky Behaviors in over 1 Million Individuals Identify Hundreds of Loci and Shared Genetic Influences." *Nature Genetics* 51 (2): 245–57.

Keller, Evelyn Fox. 2009. *The Century of the Gene*. Harvard University Press.

———. 2014. "From Gene Action to Reactive Genomes." *The Journal of Physiology* 592 (11): 2423–29.

———. 2017. "Nature, Nurture, and the Human Genome Project." In *The Ethics of Biotechnology*. Routledge.

Kuhn, T. S. 1964. "The structure of scientific revolutions" *Vol. P 159. the University of Chicago Press.*

Lakatos, Imre. 1970. "Falsification and the Methodology of Scientific Research Programmes." In *Can Theories Be Refuted?*, 205–59. Springer, Dordrecht.

Lawson, Daniel John, Neil Martin Davies, Simon Haworth, Bilal Ashraf, Laurence Howe, Andrew Crawford, Gibran Hemani, George Davey Smith, and Nicholas John Timpson. 2020. "Is Population Structure in the Genetic Biobank Era Irrelevant, a Challenge, or an Opportunity?" *Human Genetics* 139 (1): 23–41.

Layzer, D. 1974. "Heritability Analyses of IQ Scores: Science or Numerology?".*Science* **183**,1259-1266.

Lee, James J., Robbee Wedow, Aysu Okbay, Edward Kong, Omeed Maghzian, Meghan Zacher, Tuan Anh Nguyen-Viet, et al. 2018. "Gene Discovery and Polygenic Prediction from a 1.1-Million-Person GWAS of Educational Attainment." *Nature Genetics* 50 (8): 1112–21.

Levitis, Daniel A., William Z. Lidicker, and Glenn Freund. 2009. "Behavioural Biologists Do Not Agree on What Constitutes Behaviour." *Animal Behaviour* 78 (1): 103–10.

Lewis, E. B. 1978. “A Gene Complex Controlling Segmentation in Drosophila.” *Nature* 276 (5688): 565–70.

Lewontin, R C. 1974. “Annotation: The Analysis of Variance and the Analysis of Causes.” *American Journal of Human Genetics* 26 (3): 400–411.

Liu, Hexuan. 2019. “Genetic Architecture of Socioeconomic Outcomes: Educational Attainment, Occupational Status, and Wealth.” *Social Science Research* 82 (August):137–47.

Lloyd, Elisabeth A. 2015. “Adaptationism and the Logic of Research Questions: How to Think Clearly About Evolutionary Causes.” *Biological Theory* 10 (4): 343–62.

Loehlin, John C. 2009. “History of Behavior Genetics.” In *Handbook of Behavior Genetics*, edited by Yong-Kyu Kim, 3–11. New York, NY: Springer.

Longino, Helen E. 2013a. *Studying Human Behavior: How Scientists Investigate Aggression and Sexuality*. Chicago, UNITED STATES: University of Chicago Press. http://ebookcentral.proquest.com/lib/umn/detail.action?docID=1107744.

Luciano, M., D. M. Evans, N. K. Hansell, S. E. Medland, G. W. Montgomery, N. G. Martin, M. J. Wright, and T. C. Bates. 2013. “A Genome-Wide Association Study for Reading and Language Abilities in Two Population Cohorts.” *Genes, Brain and Behavior* 12 (6): 645–52.

Mackay, Trudy F. C., Stefanie L. Heinsohn, Richard F. Lyman, Amanda J. Moehring, Theodore J. Morgan, and Stephanie M. Rollmann. 2005. “Genetics and Genomics of Drosophila Mating Behavior.” *Proceedings of the National Academy of Sciences* 102 (suppl_1): 6622–29..

Maher, Brendan. 2008. “Personal Genomes: The Case of the Missing Heritability.” *Nature* 456 (7218): 18–22.

Manchia, Mirko, Stefano Comai, Martina Pinna, Federica Pinna, Vassilios Fanos, Eileen Denovan-Wright, and Bernardo Carpiniello. 2019. “Chapter Four - Biomarkers in Aggression.” In *Advances in Clinical Chemistry*, edited by Gregory S. Makowski, 93:169–237. Elsevier.

Mark, M., Rijli, F. & Chambon, P.1997. "Homeobox Genes in Embryogenesis and Pathogenesis". *Pediatric Research* **42**, 421–429.

Markow, T. A. 1987. “Behavioral and Sensory Basis of Courtship Success in Drosophila Melanogaster.” *Proceedings of the National Academy of Sciences USA* 84 (17): 6200–6204..

Martin, Nicolas W., Sarah E. Medland, Karin J. H. Verweij, S. Hong Lee, Dale R. Nyholt, Pamela A. Madden, Andrew C. Heath, Grant W. Montgomery, Margaret J. Wright, and Nicholas G. Martin. 2011. “Educational Attainment: A Genome Wide Association Study in 9538 Australians.” *PLoS ONE* 6 (6): e20128.

Matthews, Lucas J. 2020. “Precision (Mis)Education.” *Hastings Center Report* 50 (1)..

———. 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* 91:1–9.

———. 2024. “The Geneticization of Education and Its Bioethical Implications.” *Cambridge Quarterly of Healthcare Ethics*, July, 1–17.

Matthews, Lucas J., Matthew S. Lebowitz, Ruth Ottman, and Paul S. Appelbaum. 2021. “Pygmalion in the Genes? On the Potentially Negative Impacts of Polygenic Scores for Educational Attainment.” *Social Psychology of Education* 24 (3): 789–808.

Matthews, Lucas J., and Eric Turkheimer. 2021. “Across the Great Divide: Pluralism and the Hunt for Missing Heritability.” *Synthese* 198 (3): 2297–2311.

———. 2022. “Three Legs of the Missing Heritability Problem.” *Studies in History and Philosophy of Science* 93 (June):183–91.

Matthews, Lucas J., Zhijun Zhang, and Daphne O. Martschenko. 2024. “Schoolhouse Risk: *Can We Mitigate the Polygenic Pygmalion Effect?*” *Acta Psychologica* 248 (August):104403.

McGue, Matt, Yiwei Zhang, Michael B. Miller, Saonli Basu, Scott Vrieze, Brian Hicks, Steve Malone, William S. Oetting, and William G. Iacono. 2013. “A Genome-Wide Association Study of Behavioral Disinhibition.” *Behavior Genetics* 43 (5): 363–73.

Mery, Frederic, Amsale T. Belay, Anthony K.-C. So, Marla B. Sokolowski, and Tadeusz J. Kawecki. 2007. “Natural Polymorphism Affecting Learning and Memory in Drosophila.” *Proceedings of the National Academy of Sciences USA* 104 (32): 13051–55.

Miech, Richard A, and Robert M Hauser. 2001. “Socioeconomic Status and Health at Midlife: A Comparison of Educational Attainment with Occupation-Based Indicators.” *Annals of Epidemiology* 11 (2): 75–84.

Mills, Melinda C., and Charles Rahal. 2019. “A Scientometric Review of Genome-Wide Association Studies.” *Communications Biology* 2:9.

Morris, Tim T., Neil M. Davies, Gibran Hemani, and George Davey Smith. 2020. “Population Phenomena Inflate Genetic Associations of Complex Social Traits.” *Science Advances* 6 (16): eaay0328.

Mostafavi, Hakhamanesh, Arbel Harpak, Ipsita Agarwal, Dalton Conley, Jonathan K. Pritchard, and Molly Przeworski. 2020. “Variable Prediction Accuracy of Polygenic Scores within an Ancestry Group.” *eLife* 9 (January):e48376.

Nathan, Marco J., 2021. In *Black Boxes: How Science Turns Ignorance Into Knowledge*. New York, NY: Oxford University Press.

Nelson, Nicole C. 2018. *Model Behavior: Animal Experiments, Complexity, and the Genetics of Psychiatric Disorders*. University of Chicago Press.

Nelson Slater, Melissa, and Mark E. Hauber. 2017. "Olfactory Enrichment and Scent Cue Associative Learning in Captive Birds of Prey." *Zoo Biology* 36 (2): 120–26.

Nicholls, Hannah L., Christopher R. John, David S. Watson, Patricia B. Munroe, Michael R. Barnes, and Claudia P. Cabrera. 2020. "Reaching the End-Game for GWAS: Machine Learning Approaches for the Prioritization of Complex Disease Loci." *Frontiers in Genetics* 11:350.

Odintsova, Veronika V., Peter J. Roetman, Hill F. Ip, René Pool, Camiel M. Van der Laan, Klodiana-Daphne Tona, Robert R. J. M. Vermeiren, and Dorret I. Boomsma. 2019. "Genomics of Human Aggression: Current State of Genome-Wide Studies and an Automated Systematic Review Tool." *Psychiatric Genetics* 29 (5): 170–90.

Okbay, Aysu, Yeda Wu, Nancy Wang, Hariharan Jayashankar, Michael Bennett, Seyed Moeen Nehzati, Julia Sidorenko, et al. 2022. "Polygenic Prediction of Educational Attainment within and between Families from Genome-Wide Association Analyses in 3 Million Individuals." *Nature Genetics* 54 (4): 437–49.

"Orchid | Whole Genome Embryo Report." n.d. Accessed May 19, 2025. https://www.orchidhealth.com/.

Panofsky, Aaron. 2014. *Misbehaving Science: Controversy and the Development of Behavior Genetics*. Chicago ; London: The University of Chicago Press.

Parascandola, Mark. 2004. "Two Approaches to Etiology: The Debate over Smoking and Lung Cancer in the 1950s." *Endeavour* 28 (2): 81–86.

Parens, Erik, and Adrienne Ash. 2000. *Prenatal Testing and Disability Rights*. Georgetown University Press.

"Phenylketonuria (PKU) - Symptoms and Causes." n.d. Mayo Clinic. Accessed May 4, 2025. https://www.mayoclinic.org/diseases-conditions/phenylketonuria/symptoms-causes/syc-20376302.

Plaisance, K. 2006. "Behavioral Genetics and the Environment: The Generation and Exportation of Scientific Claims." Ph.D., United States -- Minnesota: University of Minnesota.

Plaisance, K., Thomas A. C. Reydon, and Mehmet Elgin. 2012. "Why the (Gene) Counting Argument Fails in the Massive Modularity Debate: The Need for Understanding Gene Concepts and Genotype-Phenotype Relationships." *Philosophical Psychology* 25 (6): 873–92.

Plaisance, K., and Thomas A. C. Reydon. 2012. "The Philosophy of Behavioral Biology." In *Philosophy of Behavioral Biology*, edited by K. Plaisance and Thomas A.C. Reydon, 3–24. Boston Studies in the Philosophy of Science. Dordrecht: Springer Netherlands.

Plomin, Robert. 2023. “Celebrating a Century of Research in Behavioral Genetics.” *Behavior Genetics* 53 (2): 75–84.

Plomin, Robert, and Sophie von Stumm. 2018. “The New Genetics of Intelligence.” *Nature Reviews Genetics* 19 (3): 148–59.

Price, Kaitlyn M., Karen G. Wigg, Else Eising, Yu Feng, Kirsten Blokland, Margaret Wilkinson, Elizabeth N. Kerr, et al. 2022. “Hypothesis-Driven Genome-Wide Association Studies Provide Novel Insights into Genetics of Reading Disabilities.” *Translational Psychiatry* 12 (1): 1–9.

Price, Kaitlyn M., Karen G. Wigg, Yu Feng, Kirsten Blokland, Margaret Wilkinson, Gengming He, Elizabeth N. Kerr, et al. 2020. “Genome-Wide Association Study of Word Reading: Overlap with Risk Genes for Neurodevelopmental Disorders.” *Genes, Brain and Behavior* 19 (6): e12648.

Rietveld, Cornelius A., Tõnu Esko, Gail Davies, Tune H. Pers, Patrick Turley, Beben Benyamin, Christopher F. Chabris, et al. 2014. “Common Genetic Variants Associated with Cognitive Performance Identified Using the Proxy-Phenotype Method.” *Proceedings of the National Academy of Sciences USA* 111 (38): 13790–94.

Rietveld, Cornelius A., Sarah E. Medland, Jaime Derringer, Jian Yang, Tõnu Esko, Nicolas W. Martin, Harm-Jan Westra, et al. 2013. “GWAS of 126,559 Individuals Identifies Genetic Variants Associated with Educational Attainment.” *Science* 340 (6139): 1467–71.

Riskin, Jessica, and M. W. Feldman. 2021. “Why Biology Is Not Destiny | M.W. Feldman and Jessica Riskin,” April 21, 2021. https://www.nybooks.com/articles/2022/04/21/why-biology-is-not-destiny-genetic-lottery-kathryn-harden/.

———. 2022. Review of *Why Biology Is Not Destiny*. The New York Review of Books, April 21, 2022. https://www.nybooks.com/articles/2022/04/21/why-biology-is-not-destiny-genetic-lottery-kathryn-harden/.

Sanders, Alan R., Gary W. Beecham, Shengru Guo, Judith A. Badner, Sven Bocklandt, Brian S. Mustanski, Dean H. Hamer, and Eden R. Martin. 2021. “Genome-Wide Linkage Study Meta-Analysis of Male Sexual Orientation.” *Archives of Sexual Behavior* 50 (8): 3371–75.

Sanders, Alan R., Gary W. Beecham, Shengru Guo, Khytam Dawood, Gerulf Rieger, Judith A. Badner, Elliot S. Gershon, et al. 2017. “Genome-Wide Association Study of Male Sexual Orientation.” *Scientific Reports* 7 (1): 16950.

Schaid, Daniel J., Wenan Chen, and Nicholas B. Larson. 2018. “From Genome-Wide Associations to Candidate Causal Variants by Statistical Fine-Mapping.” *Nature Reviews. Genetics* 19 (8): 491–504.

Scherer, Sabine, Reinhard F. Stocker, and Bertram Gerber. 2003. “Olfactory Learning in Individually Assayed Drosophila Larvae.” *Learning & Memory* 10 (3): 217–25.

Schwaerzel, Martin, Maria Monastirioti, Henrike Scholz, Florence Friggi-Grelin, Serge Birman, and Martin Heisenberg. 2003. “Dopamine and Octopamine Differentiate between Aversive and Appetitive Olfactory Memories in Drosophila.” *Journal of Neuroscience* 23 (33): 10495–502.

Scott, J.P., and J.L. Fuller. 1965. *Genetics and the Social Behavior of the Dog*. University of Chicago Press: Chicago.

Selzam, Saskia, Dale ,Philip S., Wagner ,Richard K., DeFries ,John C., Cederlöf ,Martin, O’Reilly ,Paul F., Krapohl ,Eva, and Robert and Plomin. 2017. “Genome-Wide Polygenic Scores Predict Reading Performance Throughout the School Years.” *Scientific Studies of Reading* 21 (4): 334–49.

Shero, J., W. van Dijk, A. Edwards, C. Schatschneider, E. J. Solari, and S. A. Hart. 2021. “The Practical Utility of Genetic Screening in School Settings.” *Science of Learning* 6 (1): 1–10.

Sobral, Mar. 2021. “All Traits Are Functional: An Evolutionary Viewpoint.” *Trends in Plant Science* 26 (7): 674–76.

Sokolowski, Marla B. 2001. “Drosophila: Genetics Meets Behaviour.” *Nature Reviews Genetics* 2 (11): 879–90.

Sterelny, Kim. 1988. “Critical Notice.” *Australasian Journal of Philosophy* 66 (4): 538–55.

Strisciuglio, Pietro, and Daniela Concolino. 2014. “New Strategies for the Treatment of Phenylketonuria (PKU).” *Metabolites* 4 (4): 1007–17.

Tabery, James. 2007. “Biometric and Developmental Gene–Environment Interactions: Looking Back, Moving Forward.” *Development and Psychopathology* 19 (4): 961–76.

———. 2009a. “From a Genetic Predisposition to an Interactive Predisposition: Rethinking the Ethical Implications of Screening for Gene-Environment Interactions.” *The Journal of Medicine and Philosophy: A Forum for Bioethics and Philosophy of Medicine* 34 (1): 27–48.

———. 2009b. “Making Sense of the Nature–Nurture Debate.” *Biology & Philosophy* 24 (5): 711–23.

———. 2014. *Beyond Versus: The Struggle to Understand the Interaction of Nature and Nurture*. MIT Press.

———. 2015a. “Debating Interaction: The History, and an Explanation.” *International Journal of Epidemiology* 44 (4): 1117–23.

———. 2015b. “Why Is Studying the Genetics of Intelligence So Controversial?” *Hastings Center Report* 45 (S1): S9–14.

———. 2023. *Tyranny of the Gene: Personalized Medicine and Its Threat to Public Health*. New York, UNITED STATES: Knopf Doubleday Publishing Group. http://ebookcentral.proquest.com/lib/umn/detail.action?docID=7275179.

Tolan, Patrick H., and Bennett L. Leventhal, eds. 2017. *Gene-Environment Transactions in Developmental Psychopathology*. Cham: Springer International Publishing.

Urbach, P., 1974, *Progress and Degeneration in the 'IQ Debate*'. British Journal for the Philosophy of Science, 25: 99–135, 235–259.

Valles, Sean A. 2013. "Validity and Utility in Biological Traits." *Biological Theory* 8 (1): 93–102.

Veller, Carl, and Graham M. Coop. 2024. "Interpreting Population- and Family-Based Genome-Wide Association Studies in the Presence of Confounding." *PLoS Biology* 22 (4): e3002511.

Violle, Cyrille, Marie-Laure Navas, Denis Vile, Elena Kazakou, Claire Fortunel, Irène Hummel, and Eric Garnier. 2007. "Let the Concept of Trait Be Functional!" *Oikos* 116 (5): 882–92.

Wang, Chih-Ming, Chun-Yuan Wu, Chen-En Lin, Ming-Chi Hsu, Jing-Chun Lin, Chuan-Chin Huang, Ting-Yu Lien, Hsin-Kai Lin, Ting-Wei Chang, and Hsueh-Cheng Chiang. 2023. "Forgotten Memory Storage and Retrieval in Drosophila." *Nature Communications* 14 (1): 7153.

Weir, B. S. 2008. "Linkage Disequilibrium and Association Mapping." *Annual Review of Genomics and Human Genetics* 9:129–42.

Wenzel, J. W. 1992. "Behavioral Homology and Phylogeny." *Annual Review of Ecology and Systematics* 23 (1): 361–81.

Willoughby, Emily A., Alexandros Giannelis, William G. Iacono, Matt McGue, and Scott I. Vrieze. 2022. "A Century of Behavioral Genetics at the University of Minnesota." *Twin Research and Human Genetics* 25 (6): 211–25.

Willoughby, Emily A., Alan C. Love, Matt McGue, William G. Iacono, Jack Quigley, and James J. Lee. 2019. "Free Will, Determinism, and Intuitive Judgments About the Heritability of Behavior." *Behavior Genetics* 49 (2): 136–53.

Willoughby, Emily A., Tinca J. C. Polderman, and Brian B. Boutwell. 2023. "Behavioural Genetics Methods." *Nature Reviews Methods Primers* 3 (1): 1–16. https://doi.org/10.1038/s43586-022-00191-x.

Willour, V. L., F. Seifuddin, P. B. Mahon, D. Jancic, M. Pirooznia, J. Steele, B. Schweizer, et al. 2012. "A Genome-Wide Association Study of Attempted Suicide." *Molecular Psychiatry* 17 (4): 433–44.

Wouters, Arno G. 2003. “Four Notions of Biological Function.” *Studies in History and Philosophy of Biological and Biomedical Sciences* 34 (4): 633–68.

Yamamoto, Daisuke, Jean-Marc Jallon, and Akira Komatsu. 1997. “Genetic Dissection of Sexual Behavior in Drosophila Melanogaster.” *Annual Review of Entomology* 42 (1): 551–85.

Yoshikawa, Masahiro, and Kensuke Asaba. 2021. “Educational Attainment Decreases the Risk of COVID-19 Severity in the European Population: A Two-Sample Mendelian Randomization Study.” *Frontiers in Public Health* 9 (June).

Young, Alexander I., Stefania Benonisdottir, Molly Przeworski, and Augustine Kong. 2019. “Deconstructing the Sources of Genotype-Phenotype Associations in Humans.” *Science (New York, N.Y.)* 365 (6460): 1396–1400.

Zhou, Yao, Zhiyang Zhang, Zhigui Bao, Hongbo Li, Yaqing Lyu, Yanjun Zan, Yaoyao Wu, et al. 2022. “Graph Pangenome Captures Missing Heritability and Empowers Tomato Breeding.” *Nature* 606 (7914): 527–34.

Zuk, Marlene. 2022. *Dancing Cockatoos and the Dead Man Test: How Behavior Evolves and Why It Matters*. W. W. Norton & Company.

Zuk, Marlene, and Hamish G Spencer. 2020. “Killing the Behavioral Zombie: Genes, Evolution, and Why Behavior Isn’t Special.” *BioScience* 70 (6): 515–20.

Zuk, Or, Eliana Hechter, Shamil R. Sunyaev, and Eric S. Lander. 2012. “The Mystery of Missing Heritability: Genetic Interactions Create Phantom Heritability.” *Proceedings of the National Academy of Sciences* 109 (4): 1193–98.