Chapter 3
Genome Wide Association Studies of Behavior are Social Science

Eric Turkheimer

3.1 GWAS and Its Discontents

More than a decade ago, as a half-century of population-based modeling of twin and adoption studies was giving way to the Human Genome Project and the era of measured DNA, I wrote:

> Population-based behavioral genetics has demonstrated that genotype and behavior can be expected to covary. Although the epigenetic developmental pathways linking gene products to complex behavior will in general be almost unimaginably complex, modern molecular genetics has made it possible to detect small covariations between alleles and behavior that span the complexity of the causal network….. Such associations are real and potentially interesting, but they remain correlations— and small ones— not evidence of substantial causal pathways between individual alleles and complex behavior or evidence of genes for extroversion or intelligence or evidence that future scientific efforts will be most productively applied at a genetic level of analysis. If the history of empirical psychology has taught researchers anything, it is that correlations between causally distant variables cannot be counted on to lead to coherent etiological models. (Turkheimer, 1998, p. 789)

At the time, my prediction had a distinctly Luddite ring to it. Why would anyone bet against the inexorable progress of science? My gloominess on the topic was in sharp contrast to the optimistic, not to say hegemonic, claims of most researchers at the time. Here, for example, is Plomin and Crabbe (2000) in an article entitled, "DNA": "The authors predict that in a few years, many areas of psychology will be awash in specific genes responsible for the widespread influence of genetics on behavior." (p. 806)

These predictions were made at the turn of the present century, as the Human Genome Project was realized, as human genetics made the transition from statistical

E. Turkheimer (✉)
Department of Psychology, University of Virginia, PO Box 400400, Charlottesville, VA 22904-4400, USA
e-mail: ent3c@virginia.edu

K.S. Plaisance and T.A.C. Reydon (eds.), *Philosophy of Behavioral Biology*, Boston Studies in the Philosophy of Science 282, DOI 10.1007/978-94-007-1951-4_3,

accounting of biologically related family members to the analysis of actual DNA. We are now at the end of that era, or at least it's first chapter. The technology available to genomic scientists has increased exponentially, and lately reached an apotheosis in the form of Genome Wide Association Studies, or GWAS, which allow us search through the entire genome for the bits of DNA that are more closely associated with disease or variation in normal behavior. GWAS, like so much human genomics before it, has produced somewhat paradoxical results: we are indeed, as Plomin predicted, awash in associations between human characteristics and genetic variation. At the same time, as I predicted, it is widely agreed that real scientific progress has been far more difficult than anyone expected; most, I think, would agree that new era of human genomics has been a disappointment so far. This essay will attempt to resolve this paradox, to understand how human genomics can fill libraries with "results" that nevertheless seem to fail to make progress toward the goals they were designed to reach.

3.2 Background

Genome wide association studies cannot be understood without seeing them in the historical context of behavioral genetics, which has its origins in the practical science of animal breeding. People have been breeding animals for complex characteristics, including behavioral ones, for thousands of years. The first comprehensive text about behavioral genetics, Fuller and Thomson (1960) was primarily about temperament in dogs.

Animal breeding predates both Darwin and Mendel, so much of it, whether on the farm or in the lab, was conducted without reference to anything like modern genetics. That started to change in the 20th century, although most of the traits bred in lower animals do not fit a Mendelian model of inheritance. The characteristics in Mendel's peas segregated from generation to generation: crossing wrinkled peas with smooth peas did not produce moderately wrinkled peas, but rather a mix of wrinkled and non-wrinkled, in proportions determined by the laws of classical genetics. Crossing cows high in milk production with cows low in milk production *does* produce cows with moderate levels of milk production, and selecting the highest milk producers for reproduction produces a steady increase across generations.

The classical genetics of Mendel and the genetics of complex characters like milk production was integrated (still long before anything was known about DNA) by R. A. Fisher (1918), who showed that a large number of independently segregating genes of small effect could be summed to produce a normally distributed trait that was inherited but which did not segregate. The statistical underpinning of the synthesis was based on the concept of variation. Differences among animals in milk production are associated with the degree of genetic similarity among them, as opposed to where they are raised or how they are fed, which would normally be held constant by the experimenter. The proportion of observed variation in a trait that is associated with variation in genetic relatedness is known as heritability. Heritability

is a useful concept to animal breeders, because it is related to the rate of change produced by selective breeding.

The concept of heritability can be extended to the study of humans, with some important caveats. The basis for the extension is the study of groups of people with known differences in degree of genetic relatedness, most famously identical and fraternal twins, but also siblings, parents and children, adopted (and therefore genetically unrelated) siblings, cousins, and so forth. Just as in farm animals, one can estimate a proportion of variation associated with genetic differences to the total proportion of a trait, and compute heritability between zero and one.

The crucial difference between notions of heritability in controlled studies of lower animals and studies of natural variation in humans is that for animals, the genetic and environmental variances are under the experimenter's control, and therefore fixed and uncorrelated with each other; in humans variation cannot be controlled, for obvious ethical reasons. Once variances become uncontrolled and correlated with each other, heritability coefficients no longer depend exclusively (not even primarily) on the biological characteristics of the trait in question. Instead they depend on the variability of the trait and the variation and covariation of the genes and environments that underlie it, in the particular population being studied. Having two arms notoriously has a heritability of zero, for example, because the genetic mechanisms that cause us to have two arms don't vary among individuals. Although developing two arms is intuitively and sensibly a biological process, variation in arm-number is primarily due to environmental events like accidents. One could not selectively breed cows for three-leggedness, and the reason is not that leg-number in cows is somehow essentially environmental. Rather, the genetic mechanisms involved in leg-number do not vary among cows, so it is not possible to select for them.

It is therefore not a good idea to cite heritability coefficients as a measure of how "genetic" or "environmental" something is, height included, and the high heritability of height in modern populations does not mean that it is genetically determined. One can imagine circumstances under which the heritability of height would be substantially lower (for example, under circumstances in which there were radical differences in access to adequate nutrition), and height has undergone obvious changes in recent historical time that cannot be the result of genetics. I cite the heritability of height here simply to say that height has the characteristics that lead people to think that it *ought* to be amenable to GWAS.

In any case, once the statistical means for computing the heritability of human characteristics was established, it was open season. Thousands upon thousands of family studies (mostly twin or adoption studies) were conducted, and heritabilities were computed for the usual behavioral suspects: intelligence, personality and mental illness. And to the surprise of all concerned, the studies all came out the same way: everything was heritable. Not perfectly heritable, of course, but substantially and significantly heritable. Ignoring the caveats about the interpretability of heritability coefficients in free-ranging humans, this outcome was generally taken as a great victory for genetic explanations of behavior, either to be celebrated or lamented, depending on the predisposition of the writer. I have written elsewhere (Turkheimer, 2000) about

why such conclusions turned out to be premature. The reasons can be summarized as follows, and they have resonance for the contemporary problem at hand:

1. Not only the major and established dimensions of behavior turned out to be heritable, but so did everything else. Depression is heritable, but so is marital status; intelligence is heritable, but so is how much TV people watch.
2. Heritabilities, as one might have predicted from the forgoing discussion, didn't replicate very well from study to study. They were almost never zero, but whether they were relatively high or low seemed to vary from study to study and situation to situation.
3. Largely as a consequence of (2), it is difficult to identify any major scientific advances that were produced by the twin studies, beyond the establishment that heritability is greater than zero. For example, what do we know about personality on the basis of twin studies that we did not know without them? We know that personality is moderately heritable, a fact that is not without consequences (Turkheimer, 2000), but hopes that twin studies would elucidate the causal processes underling the development of personality went mostly unfulfilled.

Such was the state of behavioral genetics at the dawn of the human genome project, which was widely viewed as a panacea for the epistemological shortcomings of twin studies. We may not have learned all that much from partitioning variance in family data, we were told, but wait until we get our hands on the actual DNA! With heritability computed in family studies as a guide (a mistaken strategy, by the way, given the inherent variability of heritability coefficients) we can now proceed to piece together the genetic processes leading to complex human traits from the ground up.

There were two main research strategies available at the outset. Linkage studies search through the genome in family pedigrees for genetic markers (locations on the genome smaller than a gene) that segregate within families in the same way as a trait of interest. Linkage studies have the advantage of being able to search the entire genome, and the disadvantage of only identifying regions, as opposed to specific locations, of interest. Association studies target specific and pre-identified genetic markers, called candidate genes, and ask whether they are correlated with the expression of a trait in the population. Association studies have the advantage of identifying relations with specific genes as opposed to regions, but are limited by our ability to decide on the candidate genes to investigate.

The newest technology, genome wide association studies, are what everyone had in mind when the genome project got underway. GWAS is the apotheosis of contemporary gene-hunting, combining many of the features of linkage and association studies. Inexpensive chips now make it easy and cheap to test for a million genetic markers in the form of single nucleotide polymorphisms, or SNPs, individual units of DNA that only take two of the four possible values of ACGT. It is thus possible to scan practically the entire genetic sequence for associations between alleles and complex traits, with a simplicity and low cost that makes it possible to include tens of thousands of research participants. Because there are so many markers across the genome, the poor focus of linkage studies has been greatly (but not completely) ameliorated, and for better or for worse one does not have to make prior identification of

the candidate genes. All that needs to be done is to find a sample of people with schizophrenia, a control group without schizophrenia, print out their genomes and look for the differences. Why wouldn't that work? But progress has been, it is safe to say, disappointing. It is not that no associations between individual alleles and specific behaviors have been found. To the contrary, we are indeed awash in them: thousands have been identified. However, the "specific" and "responsible for" clauses in Plomin and Crabbe's daring prediction have proven more difficult: despite the myriad linkages and associations between alleles and complex human traits that have been reported, three persistent limitations have proved very difficult to overcome, and they should sound familiar:

1. The reported associations are very small, in the sense that they each explain a tiny proportion of the overall variability, and collectively not much more than that;
2. The associations don't replicate very well; and
3. In part as a consequence of the first two, the various small associations between genes and behavioral outcomes haven't added up to etiological *explanations* of behaviors and especially behavioral disorders.

In other words, we are back where we started.

3.3 The Missing Heritability Problem

Others may take a rosier view than I do of the general progress that has been made toward genetic theories of behavioral syndromes, but I will save that argument for another paper. Here, I would like to discuss a remarkable series of papers published recently in *Nature Genetics*, concerning not depression or schizophrenia, not IQ or extraversion, but height. Height, that is, with near-perfect reliability of measurement and a heritability of .9 (Silventoinen et al., 2003). Height, for which there should be little problem with complex causal feedback loops. (Tall parents don't expose their children to special height-inducing environments.) Height, which has obvious biological analogs in the simplest of organisms. The genomic research paradigm, in which heritability is the gateway to identifying the specific genes composing the genetic etiology of a trait, may have turned out to be more complex than expected for juvenile delinquency, but surely it ought to work for height?

A single issue of *Nature Genetics* contained three empirical reports of genome wide association studies of height (Gudbjartsson et al., 2008; Lettre et al., 2008, Weedon et al., 2008) and a summary article describing their conclusions (Visscher, 2008). At bottom, GWAS is a search algorithm for correlations. The height studies each produced something under a half a million of them. From the outset, consideration of such results poses a problem that has been faced many times by any non-experimental social scientist: given a vast array of results that are presumably a joint reflection of some underlying process of interest, other processes of less interest that have not been controlled experimentally, and some amount of sampling error, how do you tell them apart?

The answer, of course, is null hypothesis significance testing (NHST). For any given individual association, one can compute the probability that an effect of that magnitude would occur in the sample, given a null hypothesis of no association in the population. If that probability is lower than some agreed upon "alpha" probability, one declares the null hypothesis of no association false. The alpha probability is therefore an error rate, the proportion of errors one is willing to tolerate when declaring null hypotheses false. There is, of course, another error rate involved, the "beta" or "Type-II" error rate, which describes the probability of being in error when failing to declare a null hypothesis false.

NHST is greatly complicated when there is more than one result (in this case, 400,000 results) to test. If the probability of being incorrect about any single hypothesis test is equal to α, then the probability of being incorrect on at least one of k hypothesis tests equals $1-(1-\alpha^k)$, which approaches 1.0 very quickly. Social science has developed modest technologies for dealing with the problem, like the familiar Bonferroni correction[1], but such methods do not begin to apply to the enormous number of tests conducted in GWAS, for which somewhat more sophisticated methods have been developed.

Significance testing in GWAS incorporates several steps. First, the full distribution of test probabilities is plotted against the expected distribution under the null hypothesis, to establish that *something* is disturbing the null distribution. In Weedon et al. (2008; see their Figure 1) there was an unmistakable overrepresentation of low probabilities. In the largest sample (combined meta-analytically across several studies), for example, there were 27 tests with significance levels less than 10^{-5}, compared to the four that would be expected on the basis of sampling error under the null hypothesis of no association. Weedon et al. conclude, "Approximately 23 of these loci are therefore likely to represent true positives." (p. 576)

The associations are then subjected to an even more stringent test. Thirty-nine of the original 400,000 SNPs (the 27 that exceeded the 10^{-5} criterion plus 11 that exceeded a 10^{-4} criterion, plus one more identified as a candidate in another study) were retested in an independent sample of 16,482 participants. Twenty of these 39 achieved $p<.005$ in the independent test. Combining the screening and the cross-validation, twenty SNPs had p values lower than 5×10^{-7}, 17 were lower than 10^{-8}, and 10 were lower than 10^{-10}. That's pretty significant!

But as we proceed through Weedon et al. or the other empirical reports, we find there is a second problem lurking behind the familiar one of significance testing. The statistically significant associations are further tested for something called "population stratification," and once it is found to be absent, Weedon et al. can declare, "This means that the associations are likely to reflect true biological effects on height." (p. 580) Now we would appear to be getting somewhere, although it will turn out to be problematic that no one pauses to explain what "true biological effects" are, and how they can be distinguished from biological effects that are not true or true effects that are not biological.

[1] In which the required significance level, usually $p<.05$, is divided by the total number of tests to be conducted in the experiment.

What is population stratification? The classic example of population stratification involves the discovery of a "chopsticks gene" (Hamer & Sirota, 2000). Suppose you are seeking a gene contributing to the use of chopsticks in a sample that happens to include both Asian and American participants. Any gene that differs in frequency between the Chinese and American populations will be associated with use of chopsticks, but the associations will be causally spurious. Chopstick use is *caused* by exposure to the rearing practices of Asian families; exposure to the rearing practices is *correlated with* gene frequencies, and this correlation induces a spurious one between the genes and chopstick use.

As is often the case when difficulties of this kind arise in situations where genetic methods are employed in the service of social scientific ends, the technical-sounding name that is given to the problem and to the various statistical methods that are developed to cope with it foster the impression that population stratification is essentially a technical problem in molecular genetics, to be overcome in the same way that so many other problems in genetics have been overcome, by burying them under the relentless forward momentum of contemporary genomic technology. If we can put half a million SNPs on a single chip, how big a problem can population stratification be?

But in fact, population stratification is a very old problem, and has little to do with genetics per se. Notice that population stratification doesn't arise in studies of non-human animals. That is because we have experimental control over the environments to which laboratory organisms are exposed, so we can determine that environments are either invariant or random, and there are no potential correlations between the occurrence of alleles and exposure to environments. In a horrific world in which it were possible to control the environments of humans so they could be raised identically, or randomly assigned to environments of the experimenter's choosing, population stratification would not be as severe a difficulty.

Population stratification is a problem in non-experimental causal inference, and as always, definitive attribution of causation is a matter of experimental design, not statistical analysis. A wide variety of tests, corrections and workarounds have been developed to ameliorate the effects of population stratification on GWAS. Like the original problem itself, these fixes are overlaid with a veneer of genetic technology that may lead the unwary interpreter to believe that the problem has been licked, that the science of genomic association has moved on from population stratification just as the newest SNP chip is bigger and cheaper than the last. But methodological problems in scientific inference are not so easily overcome by the next wave of technology. The fixes, moreover, are reworkings of statistical methods that have been available to social scientists for many years. And as any working social scientist is all too well aware, although the methods are sophisticated and interesting as statistical devices and useful enough as halfway measures, they don't work to discriminate true causal effects from extraneous processes that have not been controlled by the experimental method. In the long run, statistics cannot replace the causal rigor of the experimental method, no more so in genomics than in sociology.

3.4 Why not EWAS?

To a remarkable degree, GWAS was foreshadowed in a domain that might at first seem quite remote: the human social environment, and the quasi-scientific methods that have been developed to study it. The twin inferential issues in GWAS– distinguishing "true" associations from those expected on the basis of sampling error, and then distinguishing "true" causal processes from mere associations– are the bread and butter of social scientists working as far from genomics as it is possible to work. If you are a developmental psychologist trying to identify the environments that predispose some adolescents to become delinquents, what do you do? Most of the time, random assignment to environmental conditions is out of the question. So you gather as much data as possible about neighborhoods, schools, families and peers, measure delinquent outcomes in the children, and endeavor to show that some aspects of the environment *predict* (read: *are correlated with*) delinquent behavior. If you are comprehensive in your measurements of relevant environments, you might be tempted to say that you conducted an Environment Wide Association Study, or EWAS.

Of course, no self-respecting social scientist would announce such a thing because the methodological connotations are so dreadful, conjuring images of vast correlation matrices with circles around the few of them that have exceeded some magical level of statistical significance. But there is no need to be unduly derogatory about the fundamental state of affairs: in most of human behavioral science experimentation is not possible, and because it is not, scientists resort to other means, most prominent among them the analysis of systems of statistical associations. Presented with an interesting association between an environmental risk factor and a behavioral outcome, but lacking possibilities for randomized experimentation that might establish the association as causal, what would the traditional social scientist do?

The first the thing the scientist would do, of course, is to test the association for significance. For the better part of a century, far from the high-tech world of the Human Genome Project, psychologists of all persuasions have been testing their associations with NHST. From social psychologists running college students through elaborate randomized experimental conditions, to developmentalists analyzing enormous uncontrolled correlation matrices arising from observations of families, to cognitive psychologists giving repeated trials of memory tasks, to psychobiologists taking single-neuron recordings from hamster brains, to clinicians trying to establish the efficacy of psychotherapy, only two things have tied together the impossibly diverse collection of researchers that make up a psychology department: a commitment to collecting data one way or another, and an intention to test the resulting associations with NHST.

The reasons NHST has failed as a basis for scientific psychology are deep, wide, no longer a matter of serious controversy, and not the main point of this paper (see, among many others, Cohen, 1994; Schmidt, 1996). The probability levels that are computed compulsively to five decimals depend on assumptions that cannot be tested, let alone confirmed; their binary, reject or fail-to-reject formalism does violence to the subtleties of actual evaluation of scientific hypotheses in the laboratory; the tests

depend ineluctably on sample size; they encourage attention to Type I errors at the expense of attention to statistical power; the probabilities themselves represent the converse of what we really want to know, telling us the likelihood of our data given our hypothesis, when we really want the likelihood of our hypothesis being correct, given our data. These failures have been well-catalogued elsewhere and I won't do so again here (see Cohen, 1994; Harlow, Mulaik & Steiger, 1997).

In the end, the failure of NHST can be seen as a failure to solve the central dilemma of scientific psychology: for researchers working in one of the many psychological domains where randomized experimentation is impossible for practical or ethical reasons, NHST has not succeeded in discriminating actual causal processes from spurious correlations and non-causal associations. And even when experimentation *is* possible, the causal pathways leading to complex human behavior are often so diverse that empirical science seems all but helpless to unpack them, and here too NHST has provided no help.[2]

3.5 Searching for Causes in Social Science

This brings us to the next and more important, because less examined, step in the inferential chain. Given an association that passes a test of significance, how do we know if it is really causal, as opposed to the result of spurious confounds, of "population stratification"? The two broad classes of methods that are brought to bear are multivariate statistics and quasi-experimental research methods. The most basic statistical approach is multiple regression, in which possible confounds are measured and included as predictors along with the alleged causal factor. Under some restrictive conditions, the estimated regression coefficient for the factor of interest then represents its association with the outcome with values of the measured covariates "held constant" statistically. In some contexts (traditionally including situations where the effects of interest are categorical, and the potential confounds are continuous) this method is referred to as Analysis of Covariance or ANCOVA. The biggest shortcoming of multiple regression is that it requires measuring (and measuring well) all of the potential confounds of the alleged causal relationship. It is not generally possible to know if this has been accomplished successfully. Most of the multivariate alternatives to multiple regression can be characterized as attempts to circumvent the need to measure every single individual variable that might confound a causal relationship.

Principle Component Analysis, or PCA, uses the multivariate structure of the covariances among uncontrolled variables to define one or several dimensions that jointly determine the multivariate domain. So if one has measures of parental

[2]The greatest proponent of such ideas was the great theoretical psychologist Paul Meehl. The interested reader is directed to his many papers on the subject, most especially, Meehl, 1978, which should be required reading for GWAS researchers.

income, housing quality, neighborhood quality, and academic levels of local schools, one could use the positive associations among them to define a "latent variable" called *poverty*.[3] Once again under fairly restrictive assumptions, controlling for the multivariate construct succeeds in including not only the measured variables that were used to estimate it, but also the unmeasured indicators that could have been measured but weren't.

A more advanced classical method is called instrumental variable regression (Angrist, Imbens & Rubin, 1996). Given an observed association between a purported cause and an outcome, an instrument is a third variable which is correlated with the purported cause and the potential confounds, but not with the outcome, conditional on the cause and the confounds. Suppose a scientist observes an association between father-absence in families and delinquency in children: Is the relationship causal? One way to answer the question is by finding an *instrument*. In the classic example, the government might establish a new tax policy that has the effect of keeping families intact, but which would not plausibly affect rates of delinquency on its own, except by way of its correlation with intact families. Under these conditions and several other assumptions, it is possible to estimate the causal effect of intact families independent of the confounds.

A third statistical method is called propensity score analysis (Rosenbaum & Rubin, 1983). Propensity scores are a method for summarizing all of the available information about confounds of a potential cause. Returning once again to the absent father example, one way to state the problem is that because we cannot randomly assign children to absent father conditions, children with an absent father differ in many uncontrolled ways other than the father absence itself. If we collect as many possible predictors of father absence that we can think of and load them all into an equation predicting father absence, the modeled probability summarizes the overall tendency for father-present and father-absent families to be non-randomly assigned. We can match families for the overall *propensity* to have an absent father, allowing us to estimate the causal effect of absence without bias.

3.6 Within Family Designs and the Nonshared Environment

An alternative to statistical methods for establishing causation in non-experimental data is to use *quasi*-experimental designs. The range of possibilities is vast and beyond the scope of this paper (Campbell, Stanley & Gage, 1963; Rutter et al., 2001). One particular form of quasi-experimentation is particularly relevant to GWAS and

[3] A latent variable is a hypothetical process that cannot be observed directly, but which serves to explain relationships that can be observed among actual measurements. If one observes that many aspects of deprived environments—crime, poor schools, inadequate nutrition, unstimulating surroundings—tend to co-occur, the latent variable *poverty* can be invoked to explain why. The relevant statistical procedure is known as factor analysis. See MacCorquodale and Meehl (1948), or for an accessible statistical treatment, Loehlin (1992).

EWAS: within-family comparisons. Suppose you have a large sample of pairs of monozygotic (identical) twin children. Among these twins you will be able to find the occasional pair for which one member is exposed to a risk factor for delinquent behavior and the other is not. Suppose the twin who is exposed to the risk factor is indeed engaging in delinquent behavior. Is delinquency a causal consequence of the risk factor? Now at least you have an interesting control group: What is the non-exposed co-twin doing? If he is engaging in delinquent behavior to the same extent as the exposed twin, it doesn't seem likely that the risk factor *per se* is the decisive causal factor; on the other hand, if the non-exposed cotwin is not delinquent, then there may reason to expect that the risk factor *is* causing the delinquency, although as we will see below, twin designs are not capable of producing true causal inference from non-experimental data.

Within-family designs are important in many areas of psychology (Rodgers et al., 2000), and play an especially important role in behavioral genetics (Dick, Johnson, Viken & Rose, 2000), although it might be more accurate to say that within-family designs are the link between traditional behavioral genetics and the mainstream of developmental psychology. When twin studies first convinced the world of the importance of genetics in the development of human behavior (e.g., Bouchard et al., 1990), genetic variation shared supremacy with another biometric component. Although identical twins are universally more similar in behavior than fraternal twins, it is also the case that identical twins are substantially less than perfectly similar. This residual variability cannot be genetic, as identical twins are just that genetically, and it cannot be the result of differences in rearing environment, since twin pairs in these studies are raised together. The term came to be called the "nonshared environment," denoting differences among siblings or twins that arise because of environmental *differences* among children raised in the same family, as distinguished from the more intuitive "shared environment" which represents traditional socioeconomic and familial forces making family members more similar to each other. (For a philosophical treatment of the nonshared-shared environment distinction, see Plaisance, unpublished dissertation.)

In 1987, Robert Plomin and Denise Daniels published a paper with the title, "Why are Children Raised in the Same Family So Different from One Another?", in which they tried to formulate the causal processes that might underlie this variance component. Plomin and Daniels hypothesized, straightforwardly, that the characterization of the residual variance component as the nonshared environment was apt, that children raised in the same family were different from each other because their environmental experiences were different, and moreover that the specification of those differences should form the basis of environmentalist developmental psychology. They formulated a three-step program that succeeded in becoming the basis of a research program that extended over more than a decade and continues to this day:

1) Quantify the magnitude of the nonshared environmental variance component at the population level.
2) Identify environmental events that are experienced differently by children in the same family.
3) Specify the causal relations between nonshared environmental events and developmental outcomes.

In research of this kind, environmental differences between pairs of siblings or twins are used to predict differences in outcome. Perhaps most clearly in identical twin pairs, any relations that are identified cannot be attributed to genetic differences either between or within families, since the twins are genetically identical, or to environmental differences between families, like culture (chopstick use!) because the twins were raised in the same family, in the same cultural milieu. Another way of saying this is that quasi-experimental within-family designs control (imperfectly, of course) for population stratification. So the research mandated by Plomin and Daniels had two aspects that parallel the goals of contemporary GWAS. On the one hand, it was an attempt to decompose a population level variance component– the nonshared environment– into the actions of the individual environmental events it comprised; on the other, it was a quasi-experimental attempt to sift the myriad and easily-observed *associations* between environment and outcome for some smaller set that are potentially causal.

3.7 The Missing Environment Problem

In a way that once again foreshadowed the recent difficulties of the genome project, the outcome of the research mandated by Plomin and Daniels' program was disappointing. Mary Waldron and I (Turkheimer & Waldron, 2000) conducted a comprehensive meta-analysis of the research that had been conducted under the banner of the nonshared environment. In the studies we reviewed, the environment was actually measured for each member of a twin pair, rather than inferred from the twin design; just as in GWAS, DNA is now measured, as opposed to inferred from population genetics. So, for example, one might measure differences in the harshness of communications directed at siblings by their parents, and use these differences to predict differences in delinquency in the siblings. Plomin and Daniels' hypothesis can once again be stated in terms of the two aspects of the research. They hypothesized that the population-level nonshared environmental variance component could be decomposed into individual effects such as these, or equivalently, that the many non-experimental associations that are observed between risk factors and outcomes can be shown to be plausibly causal by exposing them the to within-family design.

Either way, our review demonstrated that the hypothesis could not be supported. Although the nonshared environment accounted for upwards of 50% of the variability in the studies we reviewed, the median percentage explained by any individual measured environment was under 2%. The review showed that the nonshared environmental variance component could not be decomposed into many small causal environmental events. There are substantial differences in delinquent behavior between pairs of siblings, even pairs of identical twins reared together in the same family, and the twin design can be used to establish that these differences are broadly environmental in origin. But when the investigator selects "candidate environments" that differ between siblings, for example the emotional quality of their interactions with mother, the individual effects of the candidate environments don't come close

to adding up to the total effect of "the environment" as estimated by the twin studies. Another way of saying the same thing is that observed associations between environments and outcomes—in the population, without controlling for the between-family effects of genes and shared environment, children who have more negative interactions with their mothers are more likely to be delinquent— do not stand up to the more rigorous quasi-experimental test of comparisons of siblings or twins raised together. Within families, the sibling with more negative maternal interactions is not more likely to be delinquent than the brother or sister with more positive interactions, at least not sufficiently so to account for a substantial portion of the variance component called nonshared environment. The problem of the missing variance in the nonshared environment, which was never christened as "the missing environment problem", although that is exactly what it is, remains unsolved; I remain gloomy.

The answer to the question, "Why not conduct EWAS?" is that social scientists have been conducting EWAS for 100 years. I would go so far as to assert that the history of social science before the genomic era was essentially an extended attempt at EWAS. How has it come out? The answer depends on one's opinion of the incomprehensibly large body of studies, results and evidence that environmentally-oriented social science has produced, a full evaluation of which would take us far afield. This much can be said: although environmental social science has made many interesting discoveries, and described innumerable developmental processes, some of them plausibly causal, it has not formulated comprehensive explanations of the kinds of complex human characteristics it set out to understand. There is much to learn from the thousands of environmentally-oriented studies of juvenile delinquency, divorce, depression– the list is endless– but the reader who seeks a *theory* of juvenile delinquency, or put another way, who wishes to explain, to specify, a substantial chunk of the variability in juvenile delinquency that is broadly attributed to "the environment" will not be satisfied.

There is a subtle distinction to be made here about the kinds of explanations that are possible in social science. On the one hand, to the extent the goal is to explain the environmental etiology of something like juvenile delinquency in a general sense, to identify the specific factors that cause delinquency across a broad range of contexts, only the most general, if not platitudinous, explanations can be found: poverty is bad, stable families are good. But if the question then becomes, what is it about poverty that causes delinquency, is it schooling or peer groups or diet or environmental toxins, the missing environment problem asserts itself: it is at once all of these things and none of them. Together, they all add up to the construct we call poverty, which has a demonstrably negative effect; but one at a time, their effects are too small, and too dependent on context, to be quantified reliably or added together meaningfully.

Still, the content of social science would appear to comprise more than mere repetitions of associations among generalities, although there is certainly plenty of that. Any given study of delinquency, located in a particular time and place, produces its own set of findings, in the form of particular associations among individual variables, the ones that happen to have made it over the hurdle of statistical significance in this one particular study. They may have done so simply as a result of chance, or because they really were potent causes of delinquency in the particular

socio-temporal context embodied by the sample. We usually have no way of knowing which, but either way, social science has seen so many of these significant but ephemeral associations come and go that we no longer expect very much of them.

So in social science, we have a choice. We can characterize associations among very general constructs like poverty and delinquency, which may be expected to "replicate" from one situation to the next but don't actually tell us very much about the specific causal processes that are involved. Alternatively, we can immerse ourselves in the minutiae of the particular variables that seem to be associated with delinquency in a particular time and place, which offers a satisfying sense that we are actually explaining why something happened, but frustrates us with a maddening tendency not to replicate in the next study, conducted in a subtly different context. The result is either complacent satisfaction with predictable generalities, or endless Ptolemaic theorizing about finer and finer distinctions about the outcomes of different studies, until the field gets tired of the exercise and moves on to a new phenomenon. (See Meehl's 1978 account of theorizing about the "risky shift" in the 1950s).

3.8 GWAS and EWAS

I hope that the parallels between this situation and modern genomics are now obvious. For many years in genomics, twin studies were used over and over again to re-establish the vague generality that variation in genes is correlated one way or another with variation in phenotype, with variation in *every* phenotype. After a few decades, it became clear that reasserting the heritability of something had no more actual causal content than asserting that children who live in deprived neighborhoods do worse in school, or that older children do better on developmental tests than younger children. Then modern genomics arrived, finally permitting the attempt to break down the vague concept embodied by "heritability" into the tiny molecular processes that compose it, and in the human domain we are forced to do so without the methodological advantage of randomized experimentation. The unhappy returns of GWAS are the result.

The parallel failures of EWAS and GWAS suggest that these apparent shortcomings of old-fashioned social science never did reside in the genetic naiveté of traditional environmentalists, as so many prideful behavioral geneticists have led us to believe. Instead, the problem lies in the nature of complex human behavior itself, and as such it is not really a shortcoming. We do not have a general theory of juvenile delinquency because in an important sense juvenile delinquency will not bear general theorizing. Obviously, every delinquent teenager is delinquent for some set of reasons, but the causes of one teenager's delinquency do not generalize well to the delinquency of another. (For further discussion of these ideas, see the discussion of Meehl's concept of "specific genetic etiology" in Turkheimer, 1998, and the relevant Meehl papers referenced there.)

Considering the methodological parallels between the nonshared environmental and the genomic projects promotes a humbler appreciation of the possibilities for

the latter. There is, for starters, a deep irony underlying the genome project's obsession with tiny p levels. After a century of feckless application of NHST in the face of ever-increasing philosophical and statistical condemnation of the practice, traditional social science appears finally to be giving up the ghost on significance testing. At the same time, at the outer limits of our extraordinary ability to quantify the genetic sequence, NHST is rising again. Why? Is there something about genomics that we expect to vindicate a practice discredited by half a century of unsuccessful social science?

The meager contribution of NHST to classical social science focuses our attention on exactly what is proved by the atomically small p levels achieved by the height researchers. They demonstrate, and this much we can take as conclusive notwithstanding the attendant statistical assumptions, that the observed associations between SNPs and height are very unlikely to have occurred because of sampling error. The null hypothesis that human height is unrelated to SNPs, and by extension to allelic variation, has been busted. Unfortunately, nobody ever thought such a thing in the first place, so it's a pyrrhic victory. We stand reminded: associations between SNPs and distant outcomes are associations, that is to say correlations, and absent further evidence they are nothing more than that. NHST does not provide further evidence.

So after all of the extraordinary technology of modern genomics has done its work, the study of the genetics of complex human characteristics finds itself in the same unsatisfactory scientific stance as a sociologist in 1955, trying to make sense out of a vast catalog of non-experimental survey data that purports to explain why some juveniles become delinquents while others do not. Except that the geneticist's database is even larger, and the individual associations are, if anything, smaller. The tool that is supposed to help fix things doesn't work, having been designed for the task of discriminating sampling error from population variation, rather than the identification of causal needles lost in a haystack of correlations. The tool that might actually help—randomized experimention—isn't available for ethical reasons.

In the same way, the methods of controlling for population stratification in genomics correspond point by point to the statistical and quasi-experimental methods that social scientists have been using for a century: PCA (Price et al., 2006), instrumental variables (Lawlor et al., 2008) and propensity scores (Epstein, Allen & Satten, 2007). Like their social scientific counterparts they work, more or less, but are ultimately unable to solve the broad and deep problems of causal inference that necessitated them in the first place. If a confound to an association between an allele and height is as well-behaved as the model confound of chopstick use by Asian culture, then the extant methods will identify and control for it. But what if the allele is part of a developmental process that produces a child who is more successful in demanding nutritional resources from his or her parents? Is that a height gene, a marker of a "true biological effect" on height? The variety of causal pathways that could potentially be involved in a tiny uncontrolled association is so enormous that focusing on one class of them that can be identified with some reliability borders on the futile. The point is not that the relatively small magnitude of population stratification effects should promote a sanguine view of the possibilities for raw, uncorrected GWAS, as some papers have recently suggested (Hutchison et al., 2004), but

rather that fixed statistical procedures for controlling population stratification are no more likely to correct the real problem than highly stringent significance levels.

It would be unfair not to point out that these statistical methods have some advantages when they are used in genomics, compared to their traditional use in the social sciences. The one parameter that is generally constrained by theory in twin studies- the correlation of either 1.0 or.5 between the latent genotypes of monozygotic or dizygotic twins - is exactly one parameter more than is constrained in non-genetic analyses of the same kind of behavior. The predictors, predictions, and outcomes of non-experimental social science can multiply virtually without constraint, and the modest correlational structure imposed on them by population genetic theory explains the appeal genetic modeling has for its practitioners. In addition, GWAS allows geneticists to approach an empirical standard that environmental researchers cannot match, i.e., to catalog a nearly complete record of the genetic material of individual research participants. (Contemporary methodology based on SNPs is still a step removed from the actual genetic sequence, but those remaining barriers will probably come down soon.) One reason EWAS is not possible is that the complete environmental inputs of real humans are unrecordable in principle, and also because there is no discrete environmental theory that corresponds to the intricate modern synthesis of molecular genetics, population genetics and evolutionary biology. It is hard to imagine there ever will be.

Finally, just as with the nonshared environment, within-family designs have a special place in the molecular genetics of complex phenotypes. Comparisons of parents and children or pairs of siblings offer the single most reliable way to control for population stratification. If a pair of siblings reared in the same family differs at a genetic marker and also differs in chopstick use or delinquent behavior, the association between the allelic and the behavioral differences cannot be the result of a confound resulting from exposure to different cultural environments.[4] The analogy between social scientific and genomic applications of sibling difference designs helps to show population stratification for what it is: a shared-environmental confound of an observed association. Unfortunately, the same papers that have declared population stratification a “red herring” that can safely be ignored in GWAS have specifically concluded that sib-pair analyses are too demanding (Cardon & Palmer, 2003). Collecting 65,000 individuals for a GWAS study is one thing; collecting 30,000 sibling pairs is another.

Abandoning sib-pair comparisons would be a serious error. Environmentally-oriented social science has demonstrated quite conclusively that the sibling design is a far more effective way to weed out non-experimental confounders than its statistical competitors. That so many observed associations are discounted by the sibling

[4]As was the case for within-family studies of the environment, however, the existence of within sib-pair genetic associations still do not *prove* a causal relationship between the gene and the outcome. There still might be uncontrolled confounds within pairs (one member might be sent to a Japanese school where chopstick use is encouraged, while the other goes to an American school). The within-pair association controls for a class of confounds that vary between sibling pairs, which is a big help but not a panacea for the shortcomings of non-experimental science.

comparison is not a reason to discontinue its use, but is a measure of its success. It's too bad that so many associations turn out to be non-causal when exposed to risk of disconfirmation by the within-family design, but that's the way it goes. Even the limitation on statistical power imposed by the less than astronomical size of sibling samples is probably a good thing. As the magnitude of associations gets smaller and smaller, so does the probability that we will be able to make any developmental sense out of them (Turkheimer, 2006).

3.9 Genomic Social Science and Social Scientific Genomics

At several places in this essay I have compared GWAS to something called social science. What do I mean by that? Here is a working definition: social science is a domain of inquiry into human behavior is characterized by the following:

1) There are a large number of potential causes, individually small in their effects.
2) The causes are non-independent and non-additive.
3) Randomized experimentation is not possible.

It has been widely and sometimes triumphantly noted that to remain relevant, contemporary social science must be informed by genomics and affiliated biomedical sciences like neuroanatomy and pharmacology. It is less widely recognized that the road between social science and genomics runs both ways. Old modes of explanation in the social sciences have certainly been challenged by the introduction of genetic pathways into traditional causal models, but at the same time, the glittering technologies of modern genomics are finding their limits in the centuries-old methodological complexities of human science.

The three defining characteristics of social science magnify each other in complex ways. It is not necessarily a problem, for example, that a scientific domain consists of many small causal elements. Certainly many parts of human and non-human biology are built up out of very intricate networks of small causal effects. But how are such causal processes established? They are established via randomized scientific experimentation, much of it unspeakably gruesome if breathed in the same sentence as the word "human." (William Wimsatt, 1997, tells a story of a biophysicist challenged to define his field. He said, "take an organism, homogenize it in a Waring blender, and the biophysicist is interested in those properties that are invariant under that transformation.") Much (it would be interesting to speculate about how much) of the mystery that is human behavior might be elucidated if the full experimental armamentarium of the biologist were available to the psychologist, but even considering the possibility borders on the horrific.

GWAS of complex human characteristics is social science. It is possible to conduct meaningful science under such conditions, but there are strict, and sometimes crippling, limitations on the scope of the conclusions that can be drawn. In traditional social science, successful outcomes have been produced not by mechanical application of statistical procedures to vast correlation matrices in the hope of

finding "true" effects, but rather by careful administration of quasi-experimental methods across multiple domains to detect limited instances of local regularity. This is the strategy that will be successful in human genomics as well, but it is difficult to be optimistic based on current evidence. Most GWAS research remains intent on finding "genes for" one thing or another, based on the belief that there are "true biological effects" out there to be found.

On a more optimistic note, the recent popularity of GE interaction studies represents a step in the right direction. These studies begin with one of the small associations that are detected by GWAS, and proceed to refine it by identifying environments that modify it. In the paradigmatic study of the association between a gene encoding metabolism of MAOA and antisocial behavior (Caspi et al., 2002), for example, a variant known to be associated with antisocial behavior was shown to display the effect only in the presence of a stressful rearing environment. What is interesting in terms of the argument that has been made in this paper is that such a finding represents a *restriction* on the behavioral consequences of the allele, a step back from an attempt to promulgate a general theory of the causes of violent behavior or the consequences of stressful environments or MAOA. Of such small steps successful social science is made. The extraordinary impact of this study and others like it is testimony to the need to get beyond "gene finding" and the false hope, discouraging in the long run, that genomics will bring change to the long record of slow and imperfect partial explanation in the social sciences. (For a philosophical discussion of G×E interaction, see Tabery (2009).)

3.10 Conclusion

We have yet to conclude our account of the GWAS of height. When all was said and done, across the three papers, each comprising multiple studies totaling 65,000 participants and 400,000 SNPs, assessing a trait with a heritability of.9 and a reliability of measurement greater than that, the three studies identified 20, 10 and 21 "significant" SNPs, jointly accounting for 2.9%, 2.0% and 3.7% of the total variation in height. Of the 51 SNPs identified in at least one of the three studies, eight were found in two of them, and two were found in all three. Some of the SNPs replicated those found by earlier studies, some did not; some earlier linkages were replicated, some were not.

Yet despite what one might take to be fairly discouraging results, the study authors, and especially the accompanying editorial summarizing them, adopt an upbeat and even triumphant tone. In the editorial, Visscher concluded,

> The main conclusion emerging from the current studies is that GWAS are able to robustly identify common variants that are associated with height but that the effect sizes of individual variants are small, so that very large sample sizes are needed to detect associations reliably. Single laboratories are unlikely to have sufficient sample sizes to do powerful studies on their own, and the trend in human complex trait mapping has been to create consortia of research groups and even consortia of consortia. It remains unclear at this stage how

> much genetic variation can be explained through the GWAS approach. However, if the samples in these three studies were combined together with other datasets that have been collected on height and genome-wide SNP data, then this question could be answered empirically. Genome-wide studies on, say, 100,000 individuals, unthinkable only a few years ago, will be soon be a reality. (2008, p. 490)

And what then, in the coming era of consortia of consortia? Will we be more successful in combining causally ambiguous associations each explaining a tenth of a percent of the variance than we are now when they each account for one percent?

This implacable scientific optimism has been typical of behavioral genomics since its inception. The prescribed cure for the vanishingly small effect sizes typical of genomics has always been more statistical power, in the form of ever-larger sample sizes. But at some point, the field is going to have to grapple with the possibility that the difficulty is not statistical power at all, and therefore cannot be remedied by enormous sample sizes and stringent *p* levels. No one is prone to think anymore that the answer to the environmental etiology of juvenile delinquency is to be found in larger and larger samples, allowing detection of tinier and tinier associations with environmental risks. Environmental social science has learned a bitter lesson: the explanation of behavior is difficult not because the relevant causes, though countable and essentially additive, are small and difficult to detect; rather, social science is difficult because causes are innumerable and essentially *non*-additive (Turkheimer, 2004). What causes juvenile delinquency in one place or even one person doesn't necessarily cause it in another, and whether or not a particular environmental risk causes delinquency in a particular instance depends on so many other factors, environmental and genetic, that wide-ranging scientific explanations of important phenomena are not possible.

For most complex human characteristics, the optimistically expressed but largely unexamined claims of the discovery of "true biological effects" are quixotic. Effects can be true in the sense that they have a low probability of having resulted from sampling error, as demonstrated by significance testing, but the null hypothesis that allelic variation is unrelated to complex variation is not the real issue in GWAS any more than it is in EWAS. Of course allelic variation is associated with complex outcomes: the null hypothesis is always wrong.

The claim that an effect is truly "biological" is more difficult to understand. In the limited context of population stratification, the claim presumably means that a restricted set of competing causal claims related to the actions of other alleles or environmental exposures related to them has been ruled out or corrected for, but the range of competing causal claims that might actually be made is so wide that the remediations are unconvincing and (based on evidence to date) ineffective. But in practice, the claim of a "true biological effect" is intended to connote more than a careful exclusion of a few competing causal hypotheses. The unspoken claim is that assiduous attention to statistical significance and population stratification will lead to discovery of an allele with an *identifiable biological pathway* extending through the many levels of analysis separating the allele from the complex phenomenon it is purported to explain. If I am correct that this is what the GWAS researchers intend, it is no wonder that they don't unpack the content of the claim, because on minimal

examination it is so obviously false, false even for something not-really-so-complex as height, never mind delinquency.

In the same paper that produced the quotation at the beginning of this paper (Turkheimer, 1996), I introduced a distinction between two forms of biological explanation that I called weak and strong biologism. Weak biologism is the claim, which needs nothing more than a belief in philosophical materialism to establish it, that "biology" in one form or another (usually genes or brains) underlies all complex characteristics of organisms. In the modern era, almost everyone recognizes that weak biologism is universally true: there are few vitalists or spiritualists left anymore. Weak biologism, I suggested, is why everything is heritable; it is also why everything shows a complex pattern of small associations with individual genetic markers.

Strong biologism is the claim that a complex characteristic is a consequence of a "true biological effect," the specific result of a specific event at the genomic or neurological level of analysis. The relationship between Trisomy 21 and Down Syndrome, or between a stroke lesion in the left hemisphere and a resulting aphasia, are instances of strong biologism. Strong biologism is rare and scientifically compelling. Genetically oriented behavioral scientists (in those days mostly twin researchers) I argued, had identified a fool-proof move: claim strong biological explanation on the basis of weak biological relations that depend only on the inevitable instantiation of behavior in the brain and genome.

GWAS is a reassertion of this old strategy at the molecular genetic level. The endless repetitions of genome scans that identify a few weak-to moderate signals which then don't replicate very well in the next study is simply a rediscovery on the molecular level of what I (Turkheimer, 2000) have called the First Law of Behavior Genetics: everything is heritable. Everything is heritable because of weak biologism, GWAS is always bound to produce a few "results" because everything is heritable, and heritability is instantiated in the genome, in the same not very useful sense that cognition is instantiated in the brain. The solution to the missing heritability problem is to be found in the gaps between these universal but vague concepts of physical instantiation and actual mechanistic explanation of the complex characteristics of organisms.

References

Angrist, J. D., Imbens, G. W. & Rubin, D. B. (1996): 'Identification of causal effects using instrumental variables'. *Journal of the American Statistical Association* 91: 444–455.

Bouchard, T. J., Lykken, D. T., McGue, M. Segal, N. L. & Tellegen, A. (1990): 'Sources of human psychological differences: the Minnesota study of twins reared apart'. *Science* 250: 223–228.

Campbell, D. T., Stanley, J. C. & Gage, N. L. (1963): *Experimental and Quasi-Experimental Designs for Research*. Chicago: Rand McNally.

Cardon, L. R. & Palmer, L. J. (2003): 'Populations stratification and spurious allelic association'. *The Lancet* 361: 598–604.

Caspi, A., McClay, J., Moffitt, T. E., Mill, J., Martin, J., Craig, I. W., Taylor, A., Poulton, R. (2003): 'Role of genotype in the cycle of violence in maltreated children'. *Science* 297: 851–854.

Cohen, J. (1994): 'The world is round (p<.05)'. *American Psychologist* 49: 997–1003.
Dick, D.M., Johnson, J.K, Viken, R.J. & Rose, R.J. (2000): 'Testing between-family Associations in within-family comparisons'. *Psychological Science* 11: 409–413.
Epstein, M. P., Allen, A. S., & Satten, G. A. (2007): 'A simple and improved correction for population stratification in case-control studies'. *The American Journal of Human Genetics* 80: 921–930.
Fisher, R. A. (1918): 'The correlation between relatives on the supposition of Mendelian inheritance'. *Transactions of the Royal Society of Edinburgh* 52: 399–433.
Gudbjartsson, D. F., Walters, D. F., Thorleifsson, H. S., Halldorsson, B. V., Zusmanovich, P. et al. (2008): 'Many sequence variants affecting diversity of adult human height'. *Nature Genetics* 40: 609–615.
Hamer, D. & Sirota, L. (2000): 'Beware the chopsticks gene'. *Molecular Psychiatry* 5: 11–13.
Harlow, L., Mulaik, S. A. & Steiger, J. H. (eds.) (1997): *What If There Were No Significance Tests?* Mahwah, NJ: Lawrence Erlbaum.
Hutchison, K. E., Stallings, M., McGeary, J. & Bryan, A. (2004): 'Population stratification in the candidate gene study: Fatal threat or red herring?' *Psychological Bulletin* 130: 66–79.
Lawlor, D. A., Harbord, R. M., Sterne, J. A., Timpson, N. & Davey Smith, G. (2008): 'Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology'. *Statistics in Medicine* 27: 1133–1163.
Lettre, G., Jackson, A. U., Gieger, C., Schumacher, F. R., Berndt, S. I. et al. (2008): 'Identification of ten loci associated with height highlights new biological pathways in human growth'. *Nature Genetics* 5: 584–591.
Loehlin, J. C. (1992): *Latent Variable Models: An Introduction to Factor, Path, and Structural Analysis*. Lawrence Erlbaum Associates.
MacCorquodale, K. & Meehl, P. E. (1948): 'On a distinction between hypothetical constructs and intervening variables'. *Psychological Review* 55: 95–107.
Meehl, P. E. (1978): 'Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology'. *Journal of Consulting and Clinical Psychology* 46: 806–834.
Plaisance, K. S. (2006): *Behavioral Genetics and the Environment: The Generation and Exportation of Scientific Claims*, unpublished dissertation, University of Minnesota.
Plomin, R. & Crabbe, J. (2000): 'DNA'. *Psychological Bulletin* 126: 806–828.
Plomin, R. & Daniels, D. (1987): 'Why are children in the same family so different from one another?' *Behavioral and Brain Sciences* 10: 1–16.
Price, A.L., Patterson, N.J, Plenge, R.M., Weinblatt, M.E, Shadick, N.A. & Reich, D. (2006): 'Principle components analysis corrects for stratification in genome-wide association studies'. *Nature Genetics* 38: 904–909.
Rodgers, J. L., Cleveland, H. H., van den Oord, E. & Rowe, D. C. (2000): 'Resolving the debate over birth order, family size, and intelligence'. *American Psychologist* 55: 599–612.
Rosenbaum, P. R, & Rubin, D. B. (1983): 'The central role of the propensity score in observational studies for causal effects'. *Biometrika* 70: 41–55.
Rutter, M., Pickles, A., Murray, R,. & Eaves, L. (2001): 'Testing hypotheses on specific environmental causal effects on behavior'. *Psychological Bulletin* 127: 291–324.
Schmidt, F. L. (1996): 'Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers'. *Psychological Methods* 1: 115–129.
Silventoinen et al. (2003): 'Heritability of adult body height: A comparative study of twin cohorts in eight countries'. *Twin Research and Human Genetics* 6: 399–408.
Tabery, J. (2009): 'Difference mechanisms: Explaining variation with mechanisms'. *Biology & Philosophy* 24: 645–664.
Turkheimer, E. (1998): 'Heritability and biological explanation'. *Psychological Review* 105: 782–791.
Turkheimer, E. (2000): 'Three laws of behavior genetics and what they mean'. *Current Directions in Psychological Science* 9: 160–164.
Turkheimer, E. (2004): 'Spinach and ice cream: Why social science is so difficult'. In. L. DiLalla (ed.): *Behavior Genetics Principles: Perspectives in Development, Personality, and Psychopathology*. Washington, DC, US: American Psychological Association, pp. 161–189.

Turkheimer, E. (2006): 'Interaction and play'. *PsycCRITIQUES* 51: 43.
Turkheimer, E. & Waldron, M. C. (2000): 'Nonshared environment: A theoretical, methodological and quantitative review'. *Psychological Bulletin* 126: 78–108.
Visscher, P. M. (2008): 'Sizing up human height variation'. *Nature Genetics* 40: 489–490.
Weedon, M. N., Lango, H., Lindgren, C. M., Wallace, C., Evans, D. M. et al. (2008): 'Genome-wide association analysis identifies 20 loci that influence adult height'. *Nature Genetics* 40: 575–583.
Wimsatt, W. (1997): Transcripts from "modularity of animal form". Proceedings of the evolvability of developmental mechanisms short course, http://celldynamics.org/evolvacourse/transcripts/BillWimsatt.html