Running head: CORRELATION AND CAUSATION

Correlation and Causation in the Study of Personality

James J. Lee

Department of Psychology

Harvard University

Abstract

Personality psychology aims to explain the causes and consequences of variation in behavioral traits. Because of the observational nature of the pertinent data, this endeavor has attracted much controversy. In recent years the computer scientist Judea Pearl has used a graphical approach to extend the innovations in causal inference developed by Ronald Fisher and Sewall Wright. Besides shedding much light on the philosophical notion of causality itself, this graphical framework now contains many powerful concepts of relevance to the controversies just mentioned. In this article some of these concepts are applied to areas of personality research where questions of causation arise, including the analysis of observational data and the genetic sources of individual differences.

Keywords: personality; causality; directed acyclic graph; structural equation modeling; behavioral genetics

Correlation and Causation in the Study of Personality

Consider the statement RAIN AND MUD ARE CORRELATED. Probability theory allows us to translate this bit of plain English into a mathematical language:

$$P(mud \,|\, rain) > P(mud) \quad \text{and} \quad P(rain \,|\, mud) > P(rain).$$

Translated back into words, the probability of mud increases if you have already observed rain. But what about the much stronger notion RAIN CAUSES MUD AND NOT VICE VERSA? It is surprising but true that until recently there existed no comprehensive mathematical formalism for expressing this idea. One could easily invent a new symbol—say, *do*—to indicate that the represented relation is causal and not merely correlational. Then we could write

$$P(mud \,|\, do(rain)) > P(mud) \quad \text{and} \quad P(rain \,|\, do(mud)) = P(rain).$$

to indicate that (1) rain causes mud and (2) muddying up your yard will not make it rain. Such a notational innovation is an empty gesture, however, unless it is embedded in a formal system with a rich syntax and semantics.

Unable to find a such a formal system, many scientists at the beginning of the last century dismissed causality as an ill-defined archaism. This attitude occasionally resurfaces in the literature on personality attributes such as intelligence, extraversion, political conservatism, and the like. Throughout the history of personality psychology, its practitioners have attempted to establish parts of the relational chain depicted in Figure 1. However, despite the difficulty in interpreting the chain in Figure 1 as anything but a *causal* chain, personality theorists sometimes deny that causality is within their purview (Burt, 1940; Lubinski & Dawis, 1995).

Insert Figure 1 about here

Contrary to these theorists, I take it for granted that causal knowledge is a desirable goal of the high-level sciences. In recent years the computer scientist Judea Pearl and his colleagues have greatly advanced the systematic pursuit of this goal with a formalization of causality that draws on *graph theory*. Sprites, Glymour, and Scheines (2001) and their collaborators have also made seminal contributions, although their focus is much more on the automatic generation of causal models. The graphical framework accomplishes what many Edwardian scientists thought was impossible: it captures human intuitions about causality in the form of consistent mathematical axioms. Within the structure to which these axioms give rise, one can always *prove* what can be demonstrated about causation from a given combination of data and assumptions. In this article I argue that this account of causality stands to offer a particularly great benefit to the study of personality, where for various reasons the difficulties of pursuing causal claims without a sharp causal vocabulary have been particularly keen.

Since the key mathematical objects in the graphical formalism are similar to the *path diagrams* used in *structural equation modeling* (SEM), the formalism may at first seem familiar to those scientists who already accept SEM as a technique for discerning causation in observational data. Regarding the graphical approach as an embellishment of conventional SEM practice, however, would be a mistake for at least two reasons. First, the conventional approach has been inadequately formalized and frequently abused (Freedman, 1987; McDonald & Ho, 2002), and the graphical framework supplies a necessary remedy for these shortcomings. Second, given the discipline-crossing nature of Pearl's contribution, viewing it as a refinement of a narrow and specialized and methodology would be quite blinkered. A number of commentators have emphasized that

Pearl's framework sheds philosophical light on the very notion of causality itself (Hitchcock, 2001; Gillies, 2001; Woodward, 2003).

In Part 1 I set out a relatively self-contained account of the graphical framework that will suffice for this article. Along the way I consider a problem that illustrates the graphical framework's distinctive features and is also important in its own right: what variables in a linear system must be statistically controlled to identify a causal effect using multiple regression?[1] The typical student's training may include the advice that one should control all variables that are correlated with both the putative cause and effect. This advice was criticized by Meehl (1970), and Pearl's machinery pinpoints the fallacy of this approach: there are some variables that *must* be statistically controlled and others that *must not* be so controlled. In other words it is untrue that statistically controlling another variable will either take us closer to the truth or do no harm; sometimes such "control" can take us *further* from the truth.

In Part 2 I take a necessary digression to discuss common factors—the objects of study in the psychometric tradition of personality research. A frequent objection to the scientific status of *g*, the Big Five (or Six) traits, and other factor-analytic "constructs" is that they are arbitrary mathematical fictions (Gould, 1981; Glymour, 1997). This objection is often part of a longer argument: since factor analysis is hopeless as a tool of causal discovery, any scheme that supposes common factors to be meaningful causes or consequences must be similarly unsound. Part 2 attempts to counter this nihilism. Although I also deny that a common factor is a cause of its indicators, I do allow a factor to play the role of cause or effect in graphs depicting the relations among high-level emergent entities.

Part 1 will demonstrate that any causal claim resting on observational data must at least implicitly employ SEM. Accordingly, in Part 3 I reanalyze a dataset bearing on the relation between intelligence and social liberalism in order to demonstrate how Pearl's

graphical approach can sharpen the explicit use of SEM.[2]

In Part 4 I take up the intersection of graphical methods and an emerging research area of vital importance to the entire structure depicted in Figure 1: the search for DNA polymorphisms affecting personality. The cost of sequencing a genome will eventually be negligible, and at that point gene-trait association research may succeed brain imaging as the "land grab" of behavioral science. Such research on diseases and anthropometric traits has already yielded promising dividends, including results that have been replicated across study designs, countries, and ethnicities (Teslovich et al., 2010; Waters et al., 2010; Lango Allen et al., 2010; Speliotes et al., 2010; Lanktree et al., 2011; Kooner et al., 2011; International Consortium for Blood Pressure Genome-Wide Association Studies, 2011).

Since the nature-nurture issue has been a flash point in the controversies that have dogged personality research, this article's commitment to the utility of genetic research may seem inauspicious. Here I give two related reasons for concluding my article in this way. (1) Population genetics now contains many theoretical results developed without the benefit of a general framework for causal reasoning. The new explanations of these results inspire confidence in the generality of the graphical approach. (2) Many of the examples preceding Part 4 will show that causal inferences can depend on assumptions that are untestable given the data at hand. For instance, the discussion in Part 3 invokes temporal ordering to rule out alternative models, but this assumption is admittedly fraught. A developmental process may predetermine Y well before X, even if X is measured first. Thus the soundness of any causal conclusion depends on both conforming data and the correctness of the requisite assumptions. Our substantial prior knowledge of genetics justifies many powerful assumptions, which lead to correspondingly powerful results. Gene-trait association research thus provides many enlightening applications of graphical reasoning.

Part 1: A Theory of Causality

I will now show that it is possible to state the precise conditions enabling a causal effect in a linear system to be identified using multiple regression. The preliminaries needed to formulate this important result include much of the foundations supporting Pearl's graphical framework.

Elementary Properties

Figure 2 depicts an example given by Pearl (2009, p. 15). The graph represents the causal relations among five variables: the season of the year (*season*), whether it rained last night (*rain*), whether the sprinkler was on last night (*sprinkler*), the wetness of the pavement (*wet*), and the slipperiness of the pavement (*slippery*).

Insert Figure 2 about here

The object in Figure 2 is a *directed acyclic graph* (DAG)—a collection of *nodes* and *directed edges* (single-headed arrows), each edge connecting one node to another, such that one cannot start at a node X and follow a sequence of edges along the arrows to loop back to X again. Simply put, the nodes correspond to variables and the directed edges to causal influences. The graphical framework can accommodate *cycles* representing mutual causation ($X \to Y \to \cdots \to X \to Y \to \cdots$). This paper will not address cyclic models; the reader is directed to Dickens and Flynn (2001) for an example.

In graphical parlance a *path* is a consecutive sequence of edges with distinct nodes. This terminology admittedly contradicts the SEM practice of reserving the term *path* for a single arrow between two nodes. I will conform to the convention in the broader scientific community and allow the term *path* to embrace any chain of arrows regardless of length or direction. Note that under this convention there may be more than one path connecting a

given pair of nodes. In Figure 2 both *rain* → *wet* and *rain* ← *season* → *sprinkler* → *wet* are paths between *rain* and *wet*.

If there is a directed edge from X to Y, then X is a *parent* of Y. We extend the analogy to kinship in a straightforward way to define *children*, *ancestors*, and *descendants*. This terminology enables a precise delineation of the possible reasons why two variables X and Y might be *associated* (*dependent* or *correlated*). Two reasons are well-known: (1) X is a cause of Y or vice versa, or (2) a third variable, called a *confounder*, is a common cause affecting both X and Y (Fisher, 1970).

If either X or Y is a cause of the other, then their DAG connects them with a *directed path*; each arrow along the path points in the same direction. X being a cause of Y thus corresponds, graphically, to X being an ancestor of Y. If there are any intermediate nodes between ancestor and descendant along a directed path, they are called *mediators*. In Figure 2 both *wet* → *slippery* and *season* → *rain* → *wet* are examples of directed paths; in the latter path, *rain* is a mediator.

A path in which the arrows change direction is said to be *non-directed*. The DAG representation of a confounder affecting both X and Y is a non-directed path between them that first travels against the arrows to the confounder and then travels with the arrows to terminate at the other node. In Figure 2 *rain* ← *season* → *sprinkler* supplies an example of a confounding path. *Season* is the confounder; *rain* and *sprinkler* do not affect each other, but they are associated because *season* affects both.

To better understand what directed paths mean, suppose that we wrest control of the mechanisms determining *wet* away from nature and fix the level of this variable each morning ourselves. If we use a coin flip to determine how to fix *wet* each morning, we will find that *slippery* continues to depend on *wet* but that *wet* no longer depends on *rain* or *sprinkler*. That is, if we protect the pavement with tarp whenever we are not spraying it with a garden hose, we will find that hosing the pavement is correlated with neither the

rain nor the sprinkler. The graphical representation of "overriding nature" in this way is the deletion of all directed edges converging on *wet* (Figure 2b). The intuition should be that *wet* is "set free" or "disconnected" from its parents (and other ancestors) once we intervene to determine its value. We must then attribute any persisting associations with other nodes in the graph to these nodes being descendants of *wet*. In other words a directed path encodes a persisting sensitivity of the tail node to manipulations of the head node.

Note that whether a variable is a parent (*direct cause*) or more remote ancestor (*indirect cause*) of another always depends on how deeply we understand the mechanisms at work. In Figure 2 the omission of either *rain* and *sprinkler* would force us to draw a directed edge from *season* to *wet*. That is, if we were unaware of any mediating mechanism, we would regard the time of year as directly affecting the wetness of the pavement.

Since the variables in Figure 2 are categorical, the causal relations cannot be linear. It happens that Pearl's framework is not limited to the linear models employed in many SEM applications. I will mostly restrict the discussion to linear systems for simplicity, but in the general case a node and its parents represent a variable determined by an *arbitrary* function of its direct causes.

Experimental and Statistical Control

We have just seen that experimental control amounts to physically manipulating a variable to the desired level. Can statistical control be regarded in the same way?

Recall that statistically controlling for a variable Z, in an attempt to determine whether X affects Y, amounts to observing the association between X and Y in a subpopulation where all members share the same value of Z. In the language of probability theory, we are "conditioning on" this particular value of Z. The conditional

association between X and Y will generally depend on the value assumed by Z, and ideally we would look at the relation between X and Y in each distinct subpopulation defined by a possible value of Z. However, as we condition on additional variables, the combinatorial explosion of bins defined by variable values ensures that in a small sample any particular bin contains few or no observations. For this reason we often use some kind of interpolation to predict Y from X and the *covariates* (statistically controlled variables). The simplest interpolation is the linear regression model, in which the conditional association between X and Y remains the same regardless of the covariate values. Thus, so long as a linear model is a reasonable approximation, we can speak of *the* association remaining between X and Y after conditioning on the covariates. In a linear model, "conditioning on" or "statistically controlling for" a given variable is often referred to as *partialing out* that variable. For this reason the correlation between X and Y that remains after partialing out Z is called the *partial correlation* between X and Y given Z ($\rho_{XY \cdot Z}$).

Having sorted out the terminology, let us refer back to Figure 2 to explore the consequences of statistical control. Suppose that the sprinkler has been automated such that it turns on more frequently in drier seasons. During a short timespan, rainfall will no longer be correlated with sprinkler activation. In this situation conditioning on *season* is indeed an acceptable means of determining whether there is any causal relation between *rain* and *sprinkler*. Thus, if the only non-directed paths between X and Y are confounding paths, we must statistically control a set of variables that contains at least one variable on each such path. If any association remains between X and Y, there must be at least one directed path from X to Y representing a causal effect.

Perhaps surprisingly, there are also variables that we should *not* statistically control. Earlier we named causation and confounding as two reasons for an association between variables. But there is a third reason that seems hardly known at all: X and Y may be associated because both are causes of a third variable, Z, which has been statistically

controlled. Figure 2 shows how this might occur. Although *rain* and *sprinkler* are uncorrelated if we statistically control *season*, they become correlated once again if we also statistically control *wet*. That is, if we only observe the pavement on mornings when it is wet, the two causes become *negatively* correlated; knowing that it did not rain *and* that the pavement is wet implies that the sprinkler was indeed activated.

In this situation the variable Z is a *collider*. We can think of statistically controlling a collider as *unblocking* a path between X and Y that was previously closed to causal flow. Thus, to identify the $X \to Y$ causal effect, the set of covariates must include a node on each open non-directed path between the two variables, *including* any such paths opened by conditioning on a collider or its descendants. Only then will the remaining open paths between X and Y consist solely of causal effects. If we have not conditioned on any colliders, however, we can ignore the paths including them in our attempt to estimate the $X \to Y$ causal effect.

These concepts are so crucial as to deserve their own terminology. A path between X and Y that is "closed" or "blocked" is said to be *d-separated*. A path that is not d-separated is said to *d-connect* the extreme nodes X and Y. d-separation (d-connection) is also defined for pairs of variables. Thus, a set of nodes d-separates X and Y if and only if the set blocks every path between X and Y. *Except in unusual circumstances, two variables that are d-connected must be correlated. Conversely, any two d-separated variables must be uncorrelated.*

Colliders demonstrate that *statistical* control is not equivalent to *experimental* control. Suppose that we experimentally control *wet*—again, by covering the pavement with a tarp whenever we are not spraying it with a hose. By breaking the connection between *wet* and its natural determinants (including *rain* and *sprinkler*), we are deleting the edges converging on this node (Figure 2b). This mutilation is unproblematic because the removal of edges can never add a d-connecting path. Statistically controlling the

variable, in contrast, means merely examining a subpopulation where all members happen to share the same value. Different members of this subpopulation will have that value for different reasons, which alters the covariation among the variable's causes.

The conceptual distinction between experimental and statistical control motivates Pearl's notational distinction between them. Pearl points out that when statisticians write $P(Y \mid X = x)$ to signify the (conditional) probability distribution of Y given that the variable X assumes the value x, they really mean the probability distribution of Y given that we *see* X equaling x. But what scientists want to know is the probability distribution of Y given that we *do* the action of setting X equal to x. We therefore have

$$P(Y \mid x, z) = P(Y \mid see(x), see(z)) \neq P(Y \mid do(x), see(z))$$

except in the special cases that have been described.

To show that heedless statistical control might in fact produce misleading results, I consider the model of status attainment, possibly somewhat realistic, in Figure 3. Note the use of a *bidirectional arc* to represent a dependence between two variables attributable to unmeasured common causes. In other words $X \leftrightarrow Y$ is a shorthand for $X \leftarrow C \rightarrow Y$, where C denotes the unmeasured confounders. There is some confusion in the SEM literature over the meaning of bidirectional arcs. To be clear, in the DAG approach, a bidirectional arc can *only* mean that the two variables are both affected by one or more unmeasured confounders.

For simplicity I assume that each variable in Figure 3 is well defined and measured without error. In Part 2 I will briefly comment on what these assumptions entail.

Insert Figure 3 about here

The current consensus is that we must include the directed edge *offspring IQ* $\rightarrow$

offspring SES (Murray, 2002; Nisbett, 2009). What remains under debate is the impact of IQ relative to other determinants of SES, including non-cognitive traits such as conscientiousness and agreeableness (Roberts, Kuncel, Shiner, Caspi, & Goldberg, 2007). If the SES of the parents is a confounder, the zero-order IQ-SES relation in their offspring may overestimate the causal effect of IQ. Simply including parental SES as a covariate in a regression model, however, will probably *overcorrect* the estimate. Let $C_{i,j}$ denote the unmeasured confounders represented by the bidirectional arc between nodes i and j. Statistically controlling *parent SES* d-separates the confounding paths

$$Y_4 \leftarrow Y_3 \rightarrow Y_6, \tag{1a}$$

$$Y_4 \leftarrow Y_3 \rightarrow Y_5 \rightarrow Y_6, \tag{1b}$$

$$Y_4 \leftarrow Y_3 \leftarrow Y_2 \rightarrow Y_5 \rightarrow Y_6, \tag{1c}$$

$$Y_4 \leftarrow Y_3 \leftarrow Y_2 \leftarrow C_{2,5} \rightarrow Y_5 \rightarrow Y_6, \tag{1d}$$

$$Y_4 \leftarrow C_{1,4} \rightarrow Y_1 \rightarrow Y_3 \rightarrow Y_6, \tag{1e}$$

$$Y_4 \leftarrow C_{1,4} \rightarrow Y_1 \rightarrow Y_3 \rightarrow Y_5 \rightarrow Y_6. \tag{1f}$$

Unfortunately, by unblocking the colliding paths containing $Y_1 \rightarrow Y_3 \leftarrow Y_2$, it *creates* the new d-connecting paths

$$Y_4 \leftarrow Y_1 - Y_2 \rightarrow Y_5 \rightarrow Y_6, \tag{2a}$$

$$Y_4 \leftarrow C_{1,4} \rightarrow Y_1 - Y_2 \rightarrow Y_5 \rightarrow Y_6, \tag{2b}$$

$$Y_4 \leftarrow Y_1 - Y_2 \leftarrow C_{2,5} \rightarrow Y_5 \rightarrow Y_6, \tag{2c}$$

$$Y_4 \leftarrow C_{1,4} \rightarrow Y_1 - Y_2 \leftarrow C_{2,5} \rightarrow Y_5 \rightarrow Y_6. \tag{2d}$$

The paths in (2) use an *undirected edge* between two variables to indicate that they are d-connected only after conditioning on their common descendant.

Path (2a) presents a simple case unblocking a collider by statistically controlling it. *Parent IQ* is a graphical parent of *offspring IQ*, and *parent personality trait* is a graphical

ancestor of *offspring SES*. Once our "control" of *parent SES* induces a correlation between *parent IQ* and *parent personality trait*, the flow from their nodes creates an additional d-connecting path between *offspring IQ* and *offspring SES*.

Path (2d) is instructive. Contrary to Wright's (1968) rules, this path induces a correlation despite having to go backward after already going forward. Why? After we condition on the common descendant of two causal lineages, each ancestor in one lineage will find itself d-connected with every ancestor in the other lineage. This must be true because the number of nodes in a directed path is a feature of human knowledge rather than external reality; therefore it must be possible to go from $C_{1,4}$ to $C_{2,5}$ regardless of whether any mediators along the way to the unblocked collision at *parent SES* are known. The trace goes backward from *offspring IQ* to the unobserved confounder $C_{1,4}$; this confounder is connected to $C_{2,5}$, from which the trace goes forward through *offspring personality trait* to arrive at *offspring SES*.

To summarize, the collision at *parent SES* normally impedes any causal flow through the paths in (2). Conditioning on *parent SES* unblocks the collision and allows the paths to d-connect *offspring IQ* and *offspring SES*. That is, among households *observed* to have the same SES, the covariation among the causes of SES is altered, probably becoming more negative. Whenever we have two such causes of SES, each also affecting a different member of the pair {*offspring IQ*, *offspring SES*}, they suppress the estimated magnitude of any *offspring IQ* $\rightarrow$ *offspring SES* effect. Statistically controlling for any member of {*parent IQ*, *parent personality trait*, *offspring personality trait*}, in addition to *offspring personality trait*, will restore these colliding paths to their original d-separated status. If we have not measured any of these variables, at best we can hope that the statistical control of *parent SES* removes more bias than it introduces.

The point of this exercise is not to argue for any particular model or claimed empirical finding. It is rather to demonstrate that a model-free conditioning technique,

such as the uncritical inclusion of covariates in a multiple regression, cannot be a reliable method for causal inference. The lesson is clear: *when making inferences from observational data, we should always present a DAG (structural equation model) representing our causal theory so that its critical assumptions can be criticized and defended.* In fact, one might hope that disagreements over the interpretation of observational data will often reduce to disagreements over how to connect each pair of nodes. Both sides should then find it easier to decide whether the existing data rule out any contending hypothesis and also whether any additional data can be collected to narrow the divide between them.

That said, in cases where the linearity approximation is reasonable, there is still an important role for regression in causal analysis. For instance, we may continue to encounter the naive use of multiple regression in the literature, and criteria for whether a partial regression coefficient identifies the desired causal effect are useful in judging such analyses. The following theorem sets out these criteria:

> *To identify any partial effect in a linear model, as defined by a selected set of direct or indirect paths from X to Y, we must find a set $\mathbb{S}$ of measured variables that contains no descendant of Y and d-separates all non-selected paths between X and Y. The partial effect will then equal the partial regression coefficient of X in the multiple regression of Y on $\{X\} \cup \mathbb{S}$* (Spirtes, Richardson, Meek, Scheines, & Glymour, 1998).

Whenever a report presents a partial regression coefficient as an estimate of a causal effect, one may construct plausible DAGs and determine which of these satisfy the conditions of the theorem just stated.

The Value of Randomization

Imagining the experiments implied by a DAG can sharpen our justifications for its qualitative features. Of course, the best way to ensure the feasibility of some experiment is to actually perform it.

In controlled experiments the value of the putative causal variable is assigned randomly to the participants whenever this is feasible. Why? Textbooks often invoke the fact that randomization tends to make the treatment groups well-matched on all other variables. This is a valid argument, but it may be difficult to grasp after one takes colliders into account.

The graphical framework supplies a justification of randomization that may be more intuitive. Although Fisher's (1966) argument from "the lady tasting tea" is characteristically difficult, I believe that we can rephrase it as follows. By assigning subjects to different values of a putative cause X according to a random mechanism, we are d-separating the variable from all of its ancestors. That is, since a coin flip is untouched by any arrows emanating from macroscopic variables, it follows that wiping out all arrows into X—*except* for the one coming from the coin flip—protects X from any confounders also affecting Y that may be lurking among the natural ancestors of X or the experimenter's whims. Any remaining association between X and Y then validates the causal hypothesis $X \to Y$.

Practical constraints on manipulating human circumstances may seem to render randomization a peripheral concept to personality research. In the spirit of Pearl's call to "causation without manipulation," however, we should recognize that randomization, fixing the values of confounders, and statistically controlling colliders are not the prerogatives of scientists. Nature herself engages in these activities; Part 4 will have more to say about this.

Part 2: The Nature of Psychometric Factors

Part 1 fleshed out the semantics of the verb in statements such as INTELLIGENCE CAUSES LIBERALISM. But what about the nouns in such statements?

Factor-analytic models treat measured variables, such as the different items in a personality scale, as indicators of unmeasured quantitative variables called *common factors* (Thomson, 1951; McDonald, 1985; Mulaik, 2010). In the psychometric tradition, a common factor is the generalizable quantity that any particular scale is supposed to measure imperfectly. With perhaps a tolerable loss of nuance, we can reduce questions regarding the meaningfulness of personality measurements to questions regarding the ontological status of common factors.

If the observed responses could be regressed on the unobserved factor scores, each regression coefficient would represent the quality of the scale as a measure of the corresponding factor. The regression coefficients in this model are called *factor loadings*. It follows from the regression conception that in a subpopulation where all members share the same values of a battery's common factors, the indicators making up the battery are uncorrelated. Psychometricians call this property the *principle of local independence* (Lord & Novick, 1968), and indeed some accounts begin with this principle to provide the mathematical definition of a common factor.

Any sound mathematical model must be analogous to some external reality, however, and thus the question arises: what exactly in the real world does a common factor represent? This issue has provoked recurrent debate among psychometricians. Mulaik (2005) reviews certain aspects of the controversies; noteworthy recent contributions include Borsboom, Mellenbergh, and van Heerden (2003), Molenaar (2004), Bartholomew (2004), Ashton and Lee (2005), and Bartholomew, Deary, and Lawn (2009). No writer seems to have convincingly settled the issue in a single article (or book), and I will not try to be the first. But the statement of some position, however brief and

debatable, is called for in order to move on with my attempts to employ common factors in causal explanations. In what follows I rely heavily on McDonald (1996, 2003).

Factor models are often depicted in diagrams that superficially resemble DAGs. Circles rather than boxes are used to represent common factors, and each common factor sends directed edges to the indicators measuring it. Despite the similarities, however, I maintain that the coefficients (factor loadings) attached to the edges in a factor model should *not* be interpreted as the magnitudes of causal effects. A factor model is not necessarily a causal model.

Didactic accounts of factor analysis often use the dimensions and weights of various body parts as indicators of a factor called *body size*. Now consider the proposal that body size is the unobserved cause of height, weight, and so forth. To most of us, hopefully, the notion that size causes height will seem nonsensical. An *emergent* object or property belongs to a class of phenomena that can be almost completely explained in terms of each other without reference to their low-level constituents—brain activity, cells, atoms, or whatever these constituents may be (Deutsch, 1997). Body size is a cause of those indicators that measure it, but rather is an emergent property to which the indicators are sensitive. Furthermore, a given size loading does not imply that there is some unobserved variable (but observable in principle), which, when severed from its ancestors and adjusted upward by one unit, will yield an increase in the value of the indicator equal to the loading. A large loading simply simply means that there is a high degree of conceptual overlap between the (unobservable in principle) emergent property and the (observable) indicator. Height is not the same as body size, but it is a good proxy. We might say that height makes for a passable Size Quotient.

This argument carries over to behavioral common factors. Consider the relation between extraversion and whether the respondent likes to meet new people. We can interpret the statement HE LIKES TO MEET NEW PEOPLE BECAUSE HE IS EXTRAVERTED

to mean that the respondent's behavior has an intensity that is typical of his behavior in a class of semantically related instances: whether he likes to attend parties, whether he goes out of his way to greet people, whether he enjoys public speaking, and so on. But if we construe the relation between extraversion and meeting new people as a causal one, we are saying that the respondent's behavior across a class of instances causes his behavior in a particular instance: being extraverted causes a behavior typical of an extravert. Unlike the relation between rainfall and the wetness of a pavement, the relation between extraversion and meeting new people fails to offer a means of defining the putative cause and effect independently of one another.

Someone determined to rescue the notion of a common factor as a common cause of its indicators might claim that general intelligence (g), extraversion, and other *psychometric traits* do not in fact correspond to the *folk-psychological traits* bearing these names. According to this argument, just as the physical construct of gravity bears only a metaphorical resemblance to the natural-language concept (*weight* or *seriousness*), the Big Five/Six trait of extraversion bears a resemblance of a similar kind to the natural-language concept while in fact meaning something rather different. Perhaps the simplest objection to this argument is as follows. When psychometricians want to increase the reliability of a scale, they add more indicators of the "same kind"—more items eliciting either right or wrong answers to measure intelligence, for instance, or more items inquiring about religious proclivities. This is rather telling evidence that users of factor analysis do not treat common factors as common causes. It would be a rather curious restriction on the effects of the same cause that they must all share some namable psychological-semantic property.

What about a common factor's relations to external variables? Can *these* said to be causal? For example, can body size really be said to *cause* anything? The answer to this question seems to be *yes*—if transforming someone's body so that he must be assigned a

different size factor score is a conceptually permissible manipulation. The causal claim X WON THE FIGHT BECAUSE HE IS BIGGER THAN Y then amounts to the following: if we could have fixed X's factor score to a sufficiently low value—perhaps by transplanting X's mind to a much smaller body—then X would not have prevailed over Y. Models in which other variables appear as causes of a common factor may also prove to be very useful approximations; McDonald (1996) provides the example of alcohol temporarily increasing extraversion.

In fact, if one accepts that factor analysis by itself is not a tool of causal discovery, causality only enters the picture when we consider relations with external variables. If we could complete a causal chain like the one in Figure 1, what traits would we most want to insert in the place of the node labeled *trait variation*? An evolutionary psychologist might choose those traits figuring in important theoretical accounts of human evolution. Ashton and Lee (2001) take this line in advancing their HEXACO model of personality. They have chosen a basis where three of the six axes are defined by behaviors figuring in evolutionary theories of human cooperation: Emotionality (responding to feelings of kinship and solidarity), Agreeableness (initiating exchanges, forgiving defectors), and Honesty (never defecting first, reciprocating favors). Psychologists studying other domains of individual differences might adopt this approach. Instead of attempting to find a periodic table of traits, we should try to ensure that our instruments measure traits whose causes and consequences are worth understanding. Such rationales assume the links in Figure 1 that need to be established, but surely this circularity is not a vicious one.

To summarize, common factors are personality traits that are hypothesized to exist in advance of any data analysis and can potentially be measured by an indefinite number of semantically related indicators. Such a trait is not necessarily a common cause of the indicators used to measure it, but this does not mean that the trait is a pure fiction. The adoption of psychometric methodology implies a commitment to the view that the

insertion of traits, moods, and other *intervening variables* of folk psychology between brain and behavior has proven fruitful and will continue to be necessary (MacCorquodale & Meehl, 1948).

We now have a perhaps complete taxonomy of reasons for a correlation between variables X and Y:

1. X is a cause of Y (or vice versa).
2. X and Y are both effects of a common cause.
3. X and Y are both causes of a collider that has been statistically controlled.
4. X and Y are both measures of an emergent property.

These reasons may not be mutually exclusive for a given X and Y. The last reason can never hold in the absence of at least one other.

In Part 3 I resume applications of the graphical approach, demonstrating how one can test the adequacy of the idealization entailed by employing common factors in causal explanations.

Part 3: Directed Acyclic Graphs and Structural Equation Modeling

Part 1 examined the following question: taking the system of causal relations depicted in a DAG more or less for granted, what variables must be statistically controlled to identify a linear causal effect? Here I pursue the natural followup: what assurance do we have that the DAG, as drawn, reflects reality to an acceptable degree of approximation?

The response to this vital question by orthodox SEM practitioners emphasizes the simultaneous analysis of all measured variables and global goodness-of-fit. But this approach by itself does not foreclose certain logical absurdities. For example, the measured variables may include some that are irrelevant to the important causal claims, and the contribution of such variables to the global goodness-of-fit can only obscure

judgments of model adequacy. Therefore any global fitting should be supplemented by the graphical approach advocated in this article.

Taken at face value, the orthodox view accepts the plausibility of the model

$$\Omega = \{\text{BAROMETER READINGS CAUSE RAIN}\} \cup \{\text{FRANCIS GALTON AND CHARLES DARWIN WERE COUSINS}\}.$$

When confronted with actual measurements, Ω will fit the data extremely well. The problem is that a strong correlation between certain barometer readings and rain, combined with an accurate genealogy connecting two historical figures, tells us nothing about whether barometers cause rain. We must therefore insist that the tested component of Ω (GALTON AND DARWIN WERE COUSINS) bear a logical relation to what Ω claims (THE CORRELATION BETWEEN CERTAIN BAROMETER READINGS AND RAIN MEANS THAT BAROMETERS CAUSE RAIN).

Combining the factor and causal models in one graph is a prime example of conjoining causal claims to essentially irrelevant side issues. A common procedure among personality researchers is to fit a hybrid factor-causal model and apply a rule of thumb to a scalar measure such as the goodness-of-fit index (GFI) or root mean square error of approximation (RMSEA). But if the factor model fits extremely well (and it typically will in well-motivated applications), the causal model can fit poorly without the misfit being reflected in the scalar measure. One can effect a clean divorce between measurement and causation through Anderson and Gerbing's (1988) two-step procedure: (1) test the adequacy of only the factor model, freely estimating the covariances among the factors and any non-factor variables, and then, if this step succeeds, (2) fit the causal model to the resulting covariances. Even this procedure, however, suffers from potential blurring of misfit. If there is an isolated but substantial discrepancy between the causal model and the data from step (2), adjustments in fitting other parts of that model may still produce

a scatter of small and innocent-seeming elements in the residual correlation matrix.

What is needed are *local* tests of whatever predictions are entailed by a causal model. Here is where Pearl's principle of *d*-separation becomes applicable. Recall that two variables will show a zero partial correlation once we statistically control the covariates in their *d*-separating set. A given DAG may imply certain constraints other than vanishing partial correlations; these constraints predict that a product of zero-order or partial covariances equals another such product. Whatever their form, these point predictions must hold *regardless* of the values assumed by the model parameters. Thus, to test a given DAG, we simply list the point predictions implied by a causal model and examine each one for its numerical closeness to the actual data (Shipley, 2000).

A DAG may entail many point predictions, and a problem with testing all of them is that they are not independent. For example, once the values of certain partial correlations are known, they constrain the values that other partial correlations can assume. Therefore examining every single point prediction may exaggerate the strength of the evidence for or against the hypothesized DAG. This motivates picking out a subset of the point predictions, called a *basis set*, with the following properties: (1) if all point predictions in *just* the basis set are fulfilled, then *every* point prediction implied by the DAG will also be fulfilled, and (2) no proper subset of the basis set is itself a basis set.

Breaking up a complex composite hypothesis of global fit into a basis set—a list of independently testable parts—has obvious virtues. But is it possible for this list to leave out some empirical constraints that are incorporated in the composite hypothesis? To put it differently, can a basis set miss some implications of the causal model that are in fact tested by the global fitting procedures employed in conventional SEM? The answer is *no*, as the following considerations demonstrate.

Readers familiar with the SEM notion of *covariance equivalence* will know that there may exist several distinct models that produce exactly the same fit to the covariance

matrix. A trivial example is the chain $X \rightarrow Y \rightarrow Z$, which is covariance equivalent to the reversed chain $Z \rightarrow Y \rightarrow X$ and the common-cause model $X \leftarrow Y \rightarrow Z$. Considered as DAGs, these models have the same basis set, which contains a single partial correlation: $\rho_{XZ \cdot Y}$. That is, the three models all predict that X and Z are uncorrelated after partialing out Y. The relationship between the traditional SEM notion of covariance equivalence and the graphical notion of a basis set is not an accident of this example; it is generally true that two DAGs are covariance equivalent if and only if they entail the same basis set. This graphical perspective is valuable because it provides an intuitive means of ascertaining whether two substantively contradictory models may in fact be covariance equivalent. For instance, if some alteration of a model either abolishes or introduces d-separability with respect to a pair of nodes, then the new model is not covariance equivalent to the original one. *Since models entailing the same basis set are not empirically distinguishable unless further variables are measured, a basis set exhausts* all *testable constraints that a given model imposes on a collection of measured variables.*

Note that d-separation tests of vanishing partial correlations are *not* the same as the standard SEM significance tests of estimated coefficients for at least the following three reasons. First, whereas the alternative hypothesis in d-separation is that the two nodes at issue are connected by *some* arc, the alternative hypothesis in the standard SEM approach is that the two nodes are connected by a *specific kind of arc* with a nonzero coefficient. The latter approach will produce some innocuous output even if the model has been misspecified (say by orienting the edge in the wrong direction). Second, whereas a test of vanishing partial correlation has good properties even in small samples, a standard SEM test may be valid only as the sample size becomes large.

The third distinction between d-separation and standard SEM testing depends on generalizing the notion of partial correlation to nonlinearly related variables. The standard definition of the partial correlation posits a linear system, but this restriction

can be loosened so that a partial correlation is also defined for other functional forms or even nonparametric regression techniques. As a result the qualitative correctness of a given DAG can be tested (albeit perhaps with weak power) without making any assumptions about the forms of the causal relations or the distributions of the disturbances. The standard SEM test lacks this flexibility because of its dependence on the linearity assumption.

To illustrate the graphical approach, I reanalyze a dataset presented by Deary, Batty, and Gale (2008). I follow these authors in performing separate analyses of the two sexes. Based on a sample of 3,412 males and 3,658 females, the authors concluded that a higher level of general intelligence (measured at age 11) was both a direct and indirect cause of more liberal social attitudes (measured at age 30). Figure 4 depicts their preferred model. To simplify the discussion, I retained only one of the subscales used by Deary and colleagues. Arbitrarily, I chose the subscale called *antiracism*. In the factor model, I fixed the standardized loading of the subscale on its common factor to the square root of Cronbach's α. I do not dwell on details of the factor model, which fit extremely well in both sexes.

Insert Figure 4 about here

Figure 4 belongs to a class of models whose basis sets can be characterized in the following way. For each $\{T_i, T_j\}$ not connected by neither a directed edge, consider the parents of T_i *and* the parents of T_j. The partial correlation between T_i and T_j given the *union* of their parent sets, $\rho_{ij \cdot \text{parents of } i \cup \text{parents of } j}$, must equal zero if the causal model is correct. This result should be rather intuitive; each set of parents shields its child from all d-connecting paths to the other child. Put less succinctly, if T_i and T_j do not affect each other and are not confounded by unmeasured variables, controlling for their direct causes

leaves only their probabilistically independent "error terms" to enter their partial correlation.[3]

I now proceed by finding each pair $\{T_i, T_j\}$ in Figure 4 that is not connected by an arc of any kind, since the partial correlations of these variables given their parents constitute the basis set of point predictions that I require. There are three such pairs: {*g*, *verbal residual*}, {*parent SES*, *antiracism*}, and {*SES at age 30*, *antiracism*}. Since the first pair consists of definitionally orthogonal common factors, there are only two point predictions in the basis set: after statistically controlling the parents, the partial correlations of {*parent SES*, *antiracism*} and {*SES at age 30*, *antiracism*} are equal to zero. That is, neither parental SES nor attained SES at age 30 has a direct effect on racial tolerance. At first sight this is a remarkable claim. One might have thought that changes in social circumstances might affect exposure to individuals of different backgrounds, leading in turn to changes in racial tolerance.

Insert Table 1 about here

Table 1 presents the results of the *d*-separation tests. The confidence intervals were rather wide, which shows that 4,000 participants does not approach the point of diminishing returns. Despite the ambiguities I will try to interpret the results that we have.

Since the overall model was rejected in both sexes, we are forced to a judgment of whether the numerical discrepancies were still small enough to consider the model a close approximation of reality. The partial correlation between *SES at age 30* and *antiracism* in males was the most discrepant. The sign of this partial correlation in females had the opposite sign, however, suggesting that the source of the discrepancy was small or unsystematic. Furthermore, the partial correlation between *parent SES* and *antiracism*

did indeed appear to vanish.

I have already mentioned locality as another powerful advantage of the d-separation approach. Suppose that in our judgment the partial correlation between *SES at age 30* and *antiracism* in males was too large to support their d-separability. We must then ensure that these two nodes are d-connected even after partialing out {g, *verbal residual*, *education*}. Note that insertion of the directed edge *SES at age 30* $\rightarrow$ *antiracism* will also d-connect *parent SES* and *antiracism*. If we are satisfied that these latter two nodes are d-separated by {*parent SES*, g, *verbal residual*, *education*}, we might prefer to insert the reversed edge *antiracism* $\rightarrow$ *SES at age 30*. Upon reflection this revised hypothesis is perhaps a natural one; nowadays disparaging other races may harm one's career prospects. This depth of insight into the failure of a model is typically unavailable from the modification indices provided by some software packages after an unsuccessful global fit. The statistical issues involved in "debugging" a failed model, however, require investigation.

Although the *absence* of directed edges from social status to racial tolerance an interesting finding, the primary issue in this study was the *presence* of a directed edge from g to racial tolerance. ML estimation of a linear model resulted in g showing the largest standardized direct effect on *antiracism* ($\sim$.20). But now we face a key question: what has our graphical analysis revealed so far about the trustworthiness of this estimate? If the model survives the risk posed by its basis set of point predictions ($\rho_{16 \cdot 234} = 0$ and $\rho_{56 \cdot 1234} = 0$), how much should our ensuing confidence extend to parts of the model other than the d-separable nodes?

The notion of covariance equivalence provides a ready answer: the estimate of the g $\rightarrow$ *antiracism* effect is valid if and only if there is no possible covariance-equivalent model in which this directed edge is absent, turned around, or identified in a different way. This criterion immediately reveals that many conceivable attempts to nullify the estimated g $\rightarrow$

antiracism effect can be ruled out. For example, a model that interchanges g and *antiracism* is invalid because *antiracism* can then no longer be d-separated from either *parent SES* or *SES at age 30*. Since temporal considerations weigh against most of the conceivable edge reversals, the most critical assumption is thus the absence of a bidirectional arc between g and *antiracism*. If changing $g \rightarrow$ *antiracism* to $g \leftrightarrow$ *antiracism* (or simply adding $g \leftrightarrow$ *antiracism*) preserves all vanishing partial correlations, one can place no confidence in the estimated $g \rightarrow$ *antiracism* effect. The relation between g and *antiracism* may be attributable in its entirety to confounding.

The d-separability of {*parent SES*, *antiracism*} and {*SES at age 30*, *antiracism*}, however, forbids the presence of $g \leftrightarrow$ *antiracism*. Suppose that there were such a bidirectional arc—reflecting, perhaps, a pleiotropic influence of the same genes on these two traits. Then partialing out g to d-separate {*parent SES*, *antiracism*} and {*SES at age 30*, *antiracism*} would open the colliding path *parent SES* $\leftarrow C_{1,2} - C_{2,6} \rightarrow$ *antiracism*, which could not be reblocked by any measured variable. In fact, a simple simulation shows that if there were a confounder of g and *antiracism* inducing a correlation of .20 between these two variables, partialing out g would induce a correlation of roughly $-.07$ between *parent SES* and *antiracism* that could not be removed by partialing out other variables. In summary, $g \rightarrow$ *antiracism* and $g \leftrightarrow$ *antiracism* do not predict the same vanishing partial correlations, and thus the near-zero values of the partial correlations predicted to vanish specifically under $g \rightarrow$ *antiracism* provide evidence against $g \leftrightarrow$ *antiracism*.

A similar argument shows that the $g \rightarrow$ *antiracism* estimate is robust to bidirectional arcs strongly justified by prior knowledge but which were omitted. For example, in addition to directly affecting *education* and *SES at age 30*, *parent SES* is almost certainly confounded with these two offspring characteristics. At the very least there must be personality traits, independent of abilities, that influence attainment and are themselves genetically influenced (Figure 3). Therefore these data by themselves do

not allow us to say how swapping households might have affected the attainments of this cohort. However, because the insertion of *parent SES* $\leftrightarrow$ *education* and *parent SES* $\leftrightarrow$ *SES at age 30* does not create any new d-connecting paths between g and *antiracism*, these local breakdowns of identification do not affect our estimate of the $g \rightarrow$ *antiracism* coefficient. After carrying out the d-separation tests, we can use multiple regression to estimate the coefficient of $g \rightarrow$ *antiracism* without bothering with the portions of the model that become unidentified when embedded in a more realistic supergraph.

Our conclusion is as follows. If we can somehow implement a manipulation to increase a child's level of g by age 11, it appears likely that the child will grow up to become a more racially tolerant adult. This extensive example has illustrated the distinctive features of the graphical approach to SEM, in particular highlighting how the testable implications of a causal model bear on specific substantive conclusions.

Since my reanalysis did not reach any conclusions differing from those of the original authors, the contrast between the graphical and conventional SEM approaches was not as stark as it could be. I will now recapitulate a graphical reanalysis by McDonald (2010) of an earlier SEM study to demonstrate how the conventional approach can go badly astray. The study examined five common factors: *physical health*, *daily hassles*, *world assumptions*, *constructive thinking*, and *subjective well-being*. Collectively these common factors were measured by 14 indicators. The original model posited that *physical health* and *daily hassles* affect *world assumptions* and *constructive thinking*, which in turn affect *subjective well-being*. Simplifying the history, I give credit to the original authors for recognizing that this "bottom-up" model was covariance equivalent to the "top-down" model in which *subjective well-being* is the ancestor of *physical health* and *daily hassles*.

The GFI for the global model exceeded .99. By many standards this model would be deemed acceptable. Upon fitting the factor and causal models separately, however, McDonald showed that the good global fit was attributable wholly to the good fit of the

factor model. Regardless of causal direction, ancestor and descendant must be *d*-separated by their mediators. The original model therefore predicted that the partial correlations of *subjective well-being* with both *physical health* and *daily hassles*, after statistically controlling for the intermediate variables, would equal zero. As a matter of fact, these partial correlations equaled .59 and −.12 respectively. A remarkable feature of this example is that the residual correlation matrix from the global model (of which all fit indices are a function) did not reveal any hint of where or how badly the data missed the model predictions. *Subjective well-being* must be connected to at least *physical health* with either a directed edge or a bidirectional arc, and our inability to tell these two possibilities apart means that any estimate of effects between *physical health* and *subjective well-being* may be utterly corrupted by confounding.

In this example concerning subjective well-being, the true DAG contains no *d*-separable nodes. Some commentators have argued that this DAG is representative of most interesting high-level systems (Meehl & Waller, 2002; Freedman, 2004; Greenland, 2010). Either everything affects everything, the arguments goes, or there are confounders that will never be identified. It is indeed true that for such a system "the calculation of correlation coefficients, total or partial, will not advance us a step towards evaluating the importance of the causes at work" (Fisher, 1970, p. 192). The antiracism example does suggest that the claim of ubiquitous connectedness may in fact be overly pessimistic. It is probably unwise, however, to generalize from a handful of examples. We will only know whether the causal relations within a given field are epistemologically tractable after a research effort employing the tools that have been sketched here.

Furthermore, in Part 4 I argue that there is at least one kind of causal system—the polygenic determination of a phenotype—where our prior knowledge is sufficient to dispel the intractability envisioned by skeptics of the graphical approach. *Quantitative* genetics is the branch of population genetics concerned with the genetics of continuously varying

traits (Lynch & Walsh, 1998; Bürger, 2000). Quantitative genetics has long been an integral part of personality research. It turns out that population genetics as a whole may be the basal theory needed to initiate the virtuous circle of "causal knowledge in, causal knowledge out." I now turn to the relevant aspects of this theory.

Part 4: Concepts of Genetics

Stripped of technicalities related to sample processing and delicate statistical matters, gene-trait association studies usually rely on rather simple designs: in the most straightforward case, a regression of the effect on the putative cause and a number of identically treated covariates. As I will argue, however, a replicable gene-trait *association* is nevertheless reasonably strong evidence for gene-trait *causation*. As even this modest degree of certainty is difficult to obtain in observational studies of comparable simplicity, gene hunting will be an attractive enterprise to some personality researchers seeking a foothold for the traversal of the explanatory chain in Figure 1.

I first elucidate the meaning of heritability from first principles, relying heavily on concepts that reappear in the discussion of practical issues arising in gene-trait association studies. Note that the word *gene* (or *locus*) has no single meaning. Whenever I use the term in the sense of a gene affecting a trait, I am referring to a location in a genome where discrete differences (base-pair differences, small insertions or deletions, changes in copy number, and so on) are stably inherited across generations.

Foundations of Heritability

Can one isolate, either conceptually or experimentally, the causal effects of genetic differences at a single locus? In *The Genetical Theory of Natural Selection*, Fisher introduced the concepts of the *average excess* and *average effect* to answer precisely this question. In his own words,

> Let us now consider the manner in which any quantitative individual measurement, such as human stature, may depend upon the individual genetic constitution. We may imagine, in respect of any pair of alternative [alleles], the population divided into two portions, each comprising one homozygous type together with half of the heterozygotes, which must be divided equally between the two portions. The difference in average stature between these two groups may then be termed the average excess (in stature) associated with the gene substitution in question. (Fisher, 1999, p. 30)

The average excess can be directly measured by genotyping individuals at a given locus and scoring their phenotypes.

Fisher provided two contradictory definitions of the average effect. I first consider the definition that is more suggestive of the average effect's causal meaning:

> [I]t is also necessary to give a statistical definition of a second quantity, which may be easily confused with that just defined, and may often have a nearly equal value, yet which must be distinguished from it in an accurate argument; namely the average effect produced in the population as genetically constituted, by the substitution of the one [allele] for the other. By whatever rules ... the frequency of different gene combinations, may be governed, the substitution of a small proportion of the [alleles] of one kind by the [alleles] of another will produce a definite proportional effect upon the average stature. The amount of the difference produced, on the average, in the total stature of the population, for each such gene substitution, may be termed the average effect of such substitution, in contra-distinction to the average excess as defined above. (Fisher, 1999, p. 31)

The basic notion is that a gamete is chosen at random from all those that have inherited a particular allele (say $\mathcal{A}_1$). Immediately after fertilization and before any developmental events, $\mathcal{A}_1$ is then changed to $\mathcal{A}_2$, as if by mutation. The expected change in the organism's phenotype Y at the time of measurement is then equal to the average effect. Thus, whereas all d-connecting paths between a genetic locus and the phenotype contribute to the average excess, a directed edge from the focal locus to the phenotype is necessary for a nonzero average effect. In Pearl's notation, then, the average excess is $E[(Y \mid see(\mathcal{A}_2)) - (Y \mid see(\mathcal{A}_1))]$ whereas the average effect is $E[Y \mid do(\mathcal{A}_2), see(\mathcal{A}_1)]$.

The second definition of the average effect considers a multiple regression of the trait on all loci in the genome. Now the average effect at the focal locus is equal to the partial regression coefficient of how many alleles, of the type to be counted (say $\mathcal{A}_2$), are carried by the individual (Fisher, 1941). The two definitions of the average effect agree only in special circumstances (Falconer, 1985). Because Fisher does not even mention the statistical definition based on regression in the first edition of *The Genetical Theory*, it seems that he thought the causal definition to be more fundamental, and this is how I treat it as well. There is more to be said about this; much of the next section is an argument for the pragmatic reasonableness of treating the statistical average effect as a proxy for the causal average effect.

Ignoring the distinction between the two average effects for the moment, suppose that we have a large number of loci in the genome affecting the trait Y. Let p_i be the frequency of the allele to be counted at the ith such locus. Fisher expressed the *additive genetic variance* of the trait as

$$\mathrm{Var}(A) = \sum_i 2p_i(1-p_i)a_i\alpha_i, \tag{3}$$

where a_i and α_i represent, respectively, average excess and average effect at the ith locus. The ratio of additive genetic variance to the total trait variance,

$$h^2 = \frac{\mathrm{Var}(A)}{\mathrm{Var}(Y)}, \tag{4}$$

is now known as the *heritability* of Y.

Fisher's treatment of heritability, particularly his introduction of his two averages, has struck both Price (1972) and Falconer (1985) as peculiar. It is my belief, however, that Fisher's decision in *The Genetical Theory* to base his discussion of heritability in terms of these concepts was partially motivated by his recognition of the potential for gene-trait confounding. That is, the fact that different genotypes are associated with different trait values does not by itself show that the genotypic differences *cause* the trait differences. It seems that this nicety was of great importance to Fisher. Therefore, in my recapitulation of the heritability concept, I emphasize how the distinction between confounding and causation enters into Fisher's two averages.

Geneticists refer to the confounding of genes and traits as *population structure* or *stratification*. A less formal term is the "chopstick gene syndrome": a gene showing an association with chopstick skill in a racially mixed sample is almost certainly not a gene "for" chopstick skill but rather a gene for black hair or yellow skin—or perhaps a gene where one allele has drifted by chance to high frequency in East Asians. The apocryphal story of the geneticist misled by the chopstick gene illustrates how geographical subdivision can lead to gene-trait confounding. In our evolutionary past, some humans split off from the rest of the African diaspora and became the ancestors of East Asians. Subsequently, natural selection and random genetic drift resulted in the divergence of allele frequencies among the branches of the diaspora. More recently, chopsticks were invented in China and diffused throughout what later became the Confucian belt. Thus, the ancestors of East Asians passed on both their genes and culture to their descendants,

resulting in the confounding of genotypes and chopstick skill in mixed samples of East Asians and other peoples. *Any chopstick gene will show a positive average excess in the combined mixture of subpopulations, but its average effect is in fact zero.*

Two genetic loci are in *linkage disequilibrium* (LD) if they are correlated—that is, if a person's genotype at one locus gives some information regarding the genotype at the other. This population-genetic terminology is unfortunate in that it applies even to loci not physically linked on the same chromosome, but here I abide by convention. It is important to keep in mind that a consequence of geographical subdivision is substantial LD in the global human population; that is, if a study participant carries one allele that is associated with being East Asian or some other ethnicity, it is more likely that the participant carries particular alleles associated with that ethnicity at other loci as well.

Population geneticists have shown that there are other ancestral events, including assortative mating and natural selection, that lead to LD (Fisher, 1918; Bulmer, 1971; Bürger, 2000). The mathematical soundness of these results are not in doubt, but intuitive understanding may be elusive without graphical interpretations of the kind that I now provide.

Assortative mating refers to the tendency of mated individuals to resemble each other in some phenotypic trait. Remarkably, it seems that many of us have absorbed this conspicuous fact of social life without realizing that the intuitive explanation for it (people preferring mates with certain qualities) does not correspond to anything in the canonical taxonomy of reasons for why any two variables are correlated. One mate's trait value does not affect the other mate's value, and the two trait values are not confounded in the usual sense.

The following thought experiment follows a simulation study by Eaves (1979). Although the experiment does not accurately reflect how humans mate, it does reveal how a marital correlation arising from assortative mating falls under the critical addition that

Pearl has made to the correlational taxonomy. Suppose that upon reaching a given age, all members of a cohort form random opposite-sex pairings. If the man and woman within a random couple "hit it off," they marry. The unmarried individuals may go through several more rounds of random pairing. Now suppose that after the first round we form a data matrix where each row corresponds to a randomly paired man and woman. The columns of this matrix record the trait values of each individual and also a binary variable indicating whether the two married at the end of the round. By stipulation, when considering all rows of this matrix, there is no correlation in trait value between males and females. However, if we only consider those rows where the marriage indicator assumes the value one, any traits affecting the probability of marriage become correlated. That is, marriage is a collider.

This insight into the nature of assortative mating allows us to deduce that the trait-affecting genotypes of mother and father are d-connected because of conditioning on their common effect (a successful mating). That is, those gametes carrying trait-enhancing alleles are more likely to be paired with gametes containing these same alleles. Since the paternal and maternal contributions to a recombinant gamete will both tend to contain alleles with effects on the trait of the same sign, the coupling of same-sign alleles holds within gametes as well as between them (Crow & Kimura, 1970). *All else being equal, under assortative mating the average excess will exceed the average effect; carriers of the two different alleles will tend to carry the alleles of like effect at other loci affecting the trait.*

I now turn to the confounding property of past natural selection. Fitness is a node with a multitude of directed edges converging on it from various phenotypes (Figure 5). Natural selection conditions on this node when deciding the ancestry (in the literal sense) of the offspring generation, and therefore all nodes ancestral (in the graphical sense) to fitness become d-connected. This implies that *all* functional sites in the genome are

potentially in very weak LD. In particular, if two loci affect a trait of which higher values are favored by selection, the "plus" allele at one locus is likely to be associated with the "minus" allele at the other. *Natural selection will tend to reduce the average excess below the average effect.*

Insert Figure 5 about here

I have gone through several reasons to doubt that the average excess and average effect are ever exactly equal. But under what *theoretical* circumstances, however unrealistic, do the two averages coincide? The answer to this question is insightful and also of historical interest. It can be shown that after many generations of random mating, in a broad sense that excludes not only assortative mating but natural selection and geographical subdivision, all LD and deviations from Hardy-Weinberg equilibrium will vanish (Crow & Kimura, 1970). Let us assume that there are no confounders affecting the trait through environmental mediators. Then the focal locus is d-separated from all other causes of the trait, leaving a directed edge from the locus to the phenotype as the only means by which these two nodes are connected. That is, since the two population "portions, each comprising one homozygous type together with half of the heterozygotes," do not differ in allele frequencies at any other loci, the difference in Y between them is attributable wholly to the average effect. The equivalence of the average excess and average effect under random mating is analogous to the equivalence of an observed difference and a causal effect under the randomization of treatment assignment, and indeed Fisher's (1952) thoughts on quantitative genetics stimulated his work on experimental design.

Equation 3 reveals that Fisher conceived of heritability as an inherently causal concept. Even if a locus shows a spurious average excess, its average effect must be of the

same sign for the locus to contribute to the heritability. Whenever geneticists offer a heritability estimate, then, we should interpret it as a conjecture regarding how much of the variability in the population is *caused* by genetic differences. The conjecture may be mistaken, of course, but we should separate matters of empirical adequacy from matters of definition.

Causal Inference in Gene-Trait Mapping

The correlations between the trait values of relatives are functions of the heritability and other variance components, enabling the estimation of these parameters given certain assumptions. Although some of the assumptions within a given study are approximations at best, the substantial heritabilities estimated for personality traits across different study designs nevertheless seem to justify attempts to map the DNA variants affecting these traits (Plomin, DeFries, McClearn, & McGuffin, 2008). The identification of these variants should lead to fundamental advances in our understanding of proximate mechanisms and the ultimate evolutionary forces shaping personality (Figure 1). But recall the litany of potential confounding mechanisms that may result in a divergence of the average excess (which we can directly measure) from the average effect (which we want to know).

Given the number and complexity of potential confounding mechanisms, ruling out confounding at the level of individual genetic loci may seem to pose insurmountable difficulties. The litany of confounding mechanisms, however, is actually encouraging for the following reason. Since our knowledge of the mechanisms behind confounding is typically conjectural at best, we often cannot say much about them. In contrast, the detail in which we can describe the population-genetic mechanisms behind confounding in gene-trait association studies reveals the depth of our knowledge in this domain. Exploiting our prior knowledge to characterize the relevant DAG, I argue that most sources of confounding are controllable.

Confounding cannot be a source of a gene-trait association in a family design subject to proper statistical analysis (Laird & Lange, 2011). Such designs are familiar to personality psychologists, who often study pairs of siblings reared together. There exists a positive *within-family correlation* between variables X and Y if, across sibling pairs reared together, the sibling with the higher value of X also tends to have a higher value of Y. It has been recognized that a within-family correlation presents stronger evidence for some causal relation than a correlation persisting after the statistical control of background variables (Jensen & Sinha, 1993; Turkheimer & Waldron, 2000; Beauchamp, Cesarini, Johannesson, Lindqvist, & Apicella, 2011). Pearl's distinction between *seeing* and *doing* provides a rationale for this methodological principle. Suppose that my children grow up with the family Bible always on the bookshelf. Of course, two unrelated individuals may also have grown up in households with Bibles. But whereas the chain of events depositing the Bible in the childhood home may have been quite different for each of these unrelated individuals, there is only one such chain responsible for the presence of the Bible in the home where my children will grow up. That is, within a family all background variables subsumed under "common" or "shared" environment have been *fixed* to some values, not merely *observed* to take on those values. It follows that any within-family correlation between cannot be the result of confounders that act *across* families but not *within* them.

In gene-trait association studies, an even stronger claim is justified. Mendel's Law of Segregation states that every parent possesses a pair of alleles at a given locus and passes one *randomly* selected allele to a particular offspring. The molecular basis of this law is that the reduction in meiosis of a diploid precursor cell (with two copies of each chromosome) to a haploid gamete (with one copy of each chromosome) leaves it to microlevel chance events whether any particular gamete carries any particular parental allele. Thus, when the putative cause is whether an offspring inherits $\mathcal{A}_1$ or $\mathcal{A}_2$ from a parent, treatment assignment is literally at random. Since it is nature that performs this

randomized experiment, we do not face the typical problem of deciding whether a human attempt to implement $do(x)$ is really $do(x, y, z)$.

> Genetics is indeed in a peculiarly favoured condition in that Providence has shielded the geneticist from many of the difficulties of a reliably controlled comparison. The different genotypes possible from the same mating have been beautifully randomised by the meiotic process. A more perfect control of conditions is scarcely possible, than that of different genotypes appearing in the same litter. (Fisher, 1952, p. 7)

Given a correlation between the within-family inheritance of a DNA marker and the phenotype, linkage between the marker and a causal variant is the only viable explanation. The recruitment of informative pedigrees can be difficult, however, and it is therefore desirable to seek other methods.

The fixing of genotype at fertilization restricts the class of alternative explanations for a gene-trait association. We can usually rule out reverse causation; a manipulation of a person's phenotype will typically not induce mutation. And since mutation is such a rare event, we can also discount confounding by any variable that follows fertilization in time; a confounder affecting both the DNA sequence and the trait, once development has begun, is conceivable but extremely unlikely. Given the complexity of the situation, however, this temporal restriction may initially fail to impress us. In Part 3 it was the absence of certain edges that enabled effect identification, and here we have millions of DNA sequence variants inherited from ancestors who migrated, mated, and survived natural selection in an indescribably complex way. Oddly enough, however, it turns out that this case is also conducive to effect identification. Recall that Fisher's second definition of the average effect is the partial regression coefficient of allele count in the multiple regression of the trait on all loci in the genome. The causal and statistical definitions of the average effect can coincide if gene action is purely additive, and both population-genetic theory and the

available data suggest that for many traits pure additivity should be an acceptable approximation (Hill, Goddard, & Visscher, 2008; Crow, 2010). But even if additive gene action is granted, why should the partial regression coefficient identify the causal effect? The answer comes from the graphical theorem stated in Part 1. Implicit in Fisher's second definition, then, is a claim regarding the graphical properties of gene-trait confounders.

If the ancestral confounding consists of assortative mating or natural selection in previous generations, the average excess is contaminated by confounding because of LD between the focal locus and other loci. By including all other loci in the regression, we are intercepting each and every non-directed path to the phenotype through these non-focal loci, thereby bringing the statistical and causal average effects into agreement. However, if the ancestral confounding arises from geographical subdivision or some other form of population structure, there may be non-directed paths mediated by environmental variables that have not been measured. A rather special feature of population structure allows us to overcome this difficulty in some cases: the *entire* genome is subject to the divergence of allele frequencies among subpopulations after the splintering of their ancestral population. Thus, as the number of loci entering the regression becomes very large, they become a perfect proxy for the subpopulation to which a study participant belongs. By partialing out all loci in the genome, then, we are in effect partialing out the ancestral events confounding the gene and the trait (Patterson, Price, & Reich, 2006; Kang et al., 2010).

Examples could be contrived to defeat the generalization that every confounder of gene and trait has the property of being mediated by another genetic locus or sending directed paths to an effectively infinite number of genetic loci. For example, it would not be possible for genomic background to control parental trait value directly affecting offspring trait value as in Figure 3. Nevertheless the examples of gene-trait confounding that we have examined suggest that the principle is quite robust. When combined

judiciously with family designs, studies of nominally unrelated individuals controlling for genome-wide background should be a reliable tool for pinpointing the causal effects of genetic differences. For example, the GIANT Consortium used two cohorts of families to replicate the effect sign for 150 of the 180 height-associated loci that it initially identified in studies of unrelated individuals (Lango Allen et al., 2010).

It is remarkable that observational research employing so simple a design—regression of the effect on the putative cause and a number of undifferentiated covariates—can produce such trustworthy causal inferences *in principle*. The qualifier "in principle" is necessary because of the problems introduced by *selection bias*, which occurs whenever a trait being studied is itself a cause of participation in the study. Selection bias is such an important issue, with implications extending far beyond genetics, that I will dwell on this matter in some detail.

Since an individual genetic variant is likely to have a very small effect, extremely large samples are required to detect it (Park et al., 2010). Gene hunters may have to sacrifice methodological perfectionism to attain the necessary scale. "Personal genomics" studies, drawing upon large and haphazardly ascertained all-volunteer samples, have reported associations of genetic variants with hair morphology, freckling, asparagus anosmia, photic sneeze reflex, and Parkinson's disease (Eriksson et al., 2010; Do et al., 2011). This approach will soon be extended to encompass whole-genome sequencing of similar samples exceeding 100,000 in size (Lunshof et al., 2010), and the not-too-distant future may bring even greater orders of magnitude. Now it is plausible that asparagus anosmia, say, has no effect on whether someone decides to volunteer for such a study. That is, if a person's olfactory receptors are altered in such a way that he can no longer smell the foul urine produced by an asparagus eater, the chances that the person will volunteer for a research study may well remain exactly the same. Such invariance, however, is *not* plausible for personality traits. For example, if a person's religiosity could

somehow be increased, that person may become less less inclined to participate in genetic and evolutionary research undermining his beliefs.

We can see from Figure 5 that the effect of selection bias on the divergence between the average excess and effect is qualitatively the same as that of natural selection. The quantitative effect of selection bias will typically be much stronger than that of natural selection for several reasons: (1) personality traits such as intelligence, openness, and religiosity will have much stronger effects on study participation than on fitness itself; (2) recombination has no opportunity to reduce this source of LD; and (3) any environmental effect on the trait will be negatively correlated with the number of enhancing alleles at a trait-affecting locus. The third point is not obvious, but can be understood with the aid of Figure 5. This diagram incorporates the SEM custom of using a bidirectional arc that begins and ends at the same node to represent the corresponding variable's residual disturbing causes. Explicit representation of the disturbances can greatly assist our understanding of a model, reminding us that each variable has other causes not depicted as nodes. In this case we can interpret the disturbing causes of the traits in Figure 5 as environmental in nature. Let us call the disturbance of trait 1, say, E_1, which is mnemonic for both "error" and "environment."

Even if the traits affecting study appearance are uncorrelated in the base population, these traits *and all of their causes in turn* become correlated in the selected sample as a result of the conditioning on their common effect. Thus, even if trait 2 were not at all affected by genetic variation, it would become associated with the genetic variants affecting trait 1 through the paths *gene* $\rightarrow$ *trait 1* $-$ *trait 2*. An additional problem is that the environmental causes of any particular trait are also d-connected to the causes of all other traits. For instance, suppose that a person with many "plus" alleles for religiosity volunteers for a genetic study. Then it is rather likely that the person's religiosity has been lowered by a large and negative environmental deviation, leading to no more than a

moderate level of this phenotype. Once we recognize that the disturbing causes of the traits in Figure 5 are colliding with the genetic variants, the negative correlation between genetic and environmental causes follows straightforwardly from the fact that conditioning on study participation is conditioning on a descendant of the collider (the trait). We then have the unblocked path *gene* $A - E_1 \rightarrow$ *trait 1* suppressing the estimate of the *gene* $A \rightarrow$ *trait 1* effect. The consequence of all this entanglement is reduced power to detect loci with true effects, underestimation of the average effect at any detected locus, and a surfeit of false-positive loci affecting non-focal traits that are also causes of study participation. Selection bias can even introduce bias in some family designs.

Because of the *d*-connecting paths through environmental causes, genomic background is not an adequate *d*-separating set in the presence of selection bias. It might seem from Figure 5 that we can control selection bias by including all relevant *traits* as covariates. Unfortunately this conclusion is suggested by some misleading features of this schematic DAG. First, although I have depicted the traits as causally unordered, in reality this might not be so. Those traits suspected to be causes of study participation may include colliders and mediators, and partialing out such traits invites the problems detailed at length in Part 1. Second, although for both genes and traits I have used an ellipsis to indicate that there are more nodes than depicted, a key differences is that we can sequence a whole genome but not a whole "phenome." There will surely be important causes of volunteering that we will not have measured. In contrast, the lack of a causal order among different loci in the genome and the completeness with which they can be measured is what makes genomic background such an effective shield against confounding, and we might fairly say that it is these graphical properties that gives gene-trait association studies of unrelated individuals their special character with respect to the warrant of causal inferences.

Nevertheless the measurement of those traits likely to affect appearance in a

gene-trait association study appears to be an imperfect yet desirable safeguard. Since selection bias may distort the factorial structure of personality measurements (Meredith, 1993), extra care must be taken to ensure their reliability. If a DNA marker shows an association with these traits, investigators will at least be alerted to the possibility that an additional association with some focal trait may be the result of an unblocked collision at study participation. If the association with the focal trait is the only one remaining after conditioning on the traits likely to affect study appearance, the investigators may tentatively hypothesize that the association reflects a genuine effect on the focal trait. Any firmer conclusion must await replication in a family design or a study of unrelated individuals where personal characteristics have a negligible impact on participation.

It should be clear that selection bias is a potential problem not only in genetic applications but in any observational study of socially important personality traits. The device of treating appearance in a study as a node with edges connecting it to the variables being studied can be greatly generalized to address all problems of selection bias, missing data, and unrepresentative sampling (Schafer & Graham, 2002; Little & Rubin, 2002). Some readers may be aware of the *potential-outcome* framework, which applies a taxonomy of missing-data types to these same problems. Those readers who find this framework unnatural because of its demands to consider conditional probabilities of counterfactual events may prefer the approach that I have sketched here, which requires the more intuitive judgment of whether one variable causes another. In any case Pearl (2009) has shown that the two approaches are mathematically equivalent. See Daniel, Kenward, Cousens, and De Stavola (in press) and Barenboim and Pearl (submitted) for discussion.

Conclusion

This article is in part an effort to unify the contributions of three innovators in causal reasoning: Ronald Fisher, Sewall Wright, and Judea Pearl.

Fisher began his career at a time when the distinction between correlation and causation was poorly understood and indeed scorned by leading intellectuals. Nevertheless he persisted in valuing this distinction. This led to his insight that randomization of the putative cause—whether by the deliberate introduction of "error," as his biologist colleagues thought of it, or "beautifully . . . by the meiotic process"—in fact reveals more than it obscures. His subsequent introduction of the average excess and average effect is perhaps the first explicit use of the distinction between correlation and causation in any formal scientific theory.

SEM users will know Wright—Fisher's great rival in population genetics—as the ingenious inventor of path analysis. Wright's diagrammatic approach to cause and effect serves as a conceptual bridge toward Pearl's graphical formalization, which has greatly extended the innovations developed by both of the population-genetic pioneers.

The fruitfulness of Pearl's graphical framework when applied to the problems discussed in this article bear out its utility to personality psychology. Perhaps the most surprising instance of the theory's fruitfulness concerns the role of colliders. Although obscure before Pearl's seminal work, this role turns out to be obvious in retrospect and a great aid to the understanding of covariate choice, assortative mating, selection bias, and a myriad of other seemingly unrelated problems. This article has surely only scratched the surface of the ramifications following from our recognition of colliders.

Conspicuous from these accolades by his absence is Charles Spearman—the inventor of factor analysis and thereby a founder of personality psychology. Spearman (1927) did conceive of his g factor as a hidden causal force. However, new and brilliant ideas are often only partially understood, even by their authors. After a century of theoretical

scrutiny and empirical applications, common factors appear to be more plausibly defended as mild formalizations of folk-psychological terms than as causal forces uncovered by matrix algebra. I have thus advocated a sharp distinction between the measurement of personality traits (factor analysis) and the study of their causal relations (graphical SEM). This distinction clarifies the role of factor analysis in the service that multivariate data analysis as a whole performs for personality psychology. To paraphrase McDonald (1986, p. 529), a large swath of multivariate methods can be seen as elucidating "causal relations, nonlinear in the general case, among emergent dimensions defined by indicators drawn from *a priori* behavior domains."

But this characterization bring us to a puzzle. Scientists have long used informal versions of boxes and arrows to represent hypothesized cause-effect relations. This may be because boxes and arrows effectively depict the promise of deep mechanistic understanding. Since the granularity of our boxes and arrows determines whether a given variable is a direct or indirect cause of another, it will often be possible to expand a directed edge in one graph into an entirely new subgraph. The head and tail nodes in the old graph serve as the root and sink of the new subgraph, but the two nodes are no longer in a parent-child relation. There are now intervening nodes that represent mechanisms that have been uncovered by scientific research. We can recursively continue this decomposition, substituting increasingly detailed new subgraphs for each directed edge in the graph of coarser grain. We end this recursion when each directed edge is as transparently *causal* as the collision of billiard balls or the intermeshing of gears. The wonderfully detailed illustrations of cellular processes in biology texts exemplify this level of explanation.

The puzzle is that by using common factors in our causal explanations, we seem to be retreating from this reductionistic approach. A single node called g sending an arrow to a single node called *liberalism* is surely an approximation to the true and

extraordinarily more complicated graph entangling the various physical mechanisms that underlie mental characteristics. Why this compromise? Is it sensible to test models of ethereal emergent properties shoving and being shoved by corporeal bits of matter—or, perhaps even worse, by other emergent properties? If we are committing to a calculus of causation, should we not also discard the convenient fictions of folk psychology?

The answer to this puzzle may be that reductionistic decomposition is not always the royal road to scientific understanding.

> [T]he structure of scientific explanation does not reflect the reductionistic hierarchy. Many of them are autonomous, referring only to concepts at that particular level (for instance, 'the bear ate the honey because it was hungry'). Many involve deductions in the opposite direction to that of reductive explanation. That is, they explain things not by analysing them into smaller, simpler things but by regarding them as components of larger, more complex things—about which we nevertheless have explanatory theories. For example, consider one particular copper atom at the tip of the nose of the statue of Sir Winston Churchill that stands in Parliament Square in London. Let me try to explain why that copper atom is there. It is because Churchill served as prime minister in the House of Commons nearby; and because his ideas and leadership contributed to the Allied victory in the Second World War; and because it is customary to honour such people by putting up statues of them; and because bronze, a traditional material for such statues, contains copper, and so on. Thus we explain a low-level physical observation—the presence of a copper atom at a particular location—through extremely high-level theories about emergent phenomena such as ideas, leadership, war and tradition.
>
> There is no reason why there should exist, even in principle, any lower-level *explanation* of the presence of that copper atom than the one I have just given.

> Presumably a reductive 'theory of everything' would in principle make a low-level *prediction* of the probability that such a statue will exist, given the condition of (say) the solar system at some earlier date. It would also in principle describe how the statue probably got there. But such descriptions and predictions (wildly infeasible, of course) would explain nothing. They would merely describe the trajectory that each copper atom followed from the copper mine, through the smelter and the sculptor's studio, and so on.... In fact such a prediction would have to refer to atoms all over the planet, engaged in the complex motion we call the Second World War, among other things. But even if you had the superhuman capacity to follow such lengthy predictions of the copper atom's being there, you would still not be able to say, 'Ah yes, now I understand why it is there....' You would have to inquire into what it was about that configuration of atoms, and those trajectories, that gave them the propensity to deposit a copper atom at this location. Pursuing this inquiry would be a creative task, as discovering new explanations always is. You would have to discover that certain atomic configurations support emergent phenomena such as leadership and war, which are related to one another by high-level explanatory theories. Only when you knew those theories could you understand fully why that copper atom is where it is. (Deutsch, 1997, pp. 21–23)

I find this passage persuasive. When we seek to explain high-level phenomena, we must avoid the error, criticized by Deutsch, of *vulgar reductionism*. This is the attitude that *all* legitimate scientific explanations must break up high-level phenomena into lower-level constituents. We must also avoid the converse error of *vulgar holism*, which posits that all legitimate scientific explanations ignore fundamental constituents and focus exclusively on emergent properties. In fact, we already have at least one good example of a science with

a blend of reductionistic and holistic explanations. It is surely not a coincidence that Fisher and Wright were both among the founders of population genetics in addition to being innovators in causal reasoning. Evolutionary biology is already rich and autonomous without reducing the ideas of *genotype*, *phenotype*, *fitness*, *selection*, and *adaptation* to microlevel bits and pieces. It is natural for the notion for causality to have been developed by evolutionists because the (high-level) distinction between correlation and causation, while being tangential to much of the older physical sciences, lies at the core of the evolutionary ideas just mentioned.

What kind of rich and autonomous theoretical structure, blending reductionistic and holistic elements, will emerge from the interdisciplinary field of personality psychology? Given the interests of many personality psychologists in genetics and evolution (Ashton & Lee, 2001; Penke, Denissen, & Miller, 2007; Johnson, Penke, & Spinath, 2011), a mature science of personality might inherit some of its explanatory structure from neo-Darwinism. But Figure 1 shows that there is much else about personality to be explained. Even after a century it is still difficult to offer any global perspective that can claim to be more than an opinion. The challenges of the field are daunting, and progress is gradual. But because common factors (folk-psychological traits) are the product of a cognitive and historical process that seems quite efficient at extracting powerful compressions of reality (Baum, 2004; Ashton & Lee, 2005), I suspect that they will continue to play some role within the personality psychology of the future.

Whatever the fate of common factors in causal theories, a perhaps more fundamental question is whether we can reason precisely about causality itself. An affirmative answer is the central message of this article:

> It is true that testing for cause and effect is difficult. Discovering causes of effects is even more difficult. But causality is not *mystical* or *metaphysical*.... [I]t can be expressed in a friendly mathematical language, ready for computer

analysis.

What I have presented to you . . . is a sort of pocket calculator, an *abacus*, to help us investigate certain problems of cause and effect with mathematical precision. This does not solve all the problems of causality, but the power of *symbols* and mathematics should not be underestimated. . . .

[T]he really challenging problems lie ahead: We still do not have a causal understanding of *poverty* and *cancer* and *intolerance*, and only the accumulation of data and the insight of great minds will lead to such understanding.

The data is all over the place, the insight is yours, and now an abacus is at your disposal, too. (Pearl, 2009, pp. 427–428)

References

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin*, *103*, 411–423.

Ashton, M. C., & Lee, K. (2001). A theoretical basis for the major dimensions of personality. *European Journal of Personality*, *15*, 327–353.

Ashton, M. C., & Lee, K. (2005). A defence of the lexical approach to the study of personality structure. *European Journal of Personality*, *19*, 5–24.

Barenboim, E., & Pearl, J. (submitted). *Controlling selection bias in causal inference.*

Bartholomew, D. J. (2004). *Measuring intelligence: Facts and fallacies.* Cambridge, UK: Cambridge University Press.

Bartholomew, D. J., Deary, I. J., & Lawn, M. (2009). A new lease of life for Thomson's bonds model of intelligence. *Psychological Review*, *116*, 567–579.

Baum, E. B. (2004). *What is thought?* Cambridge, MA: MIT Press.

Beauchamp, J. P., Cesarini, D., Johannesson, M., Lindqvist, E., & Apicella, C. (2011). On the sources of the height-intelligence correlation: New insights from a bivariate ACE model with assortative mating. *Behavior Genetics*, *41*, 242–252.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review*, *110*, 203–218.

Bulmer, M. G. (1971). The effect of selection on genetic variability. *American Naturalist*, *105*, 201–211.

Bürger, R. (2000). *The mathematical theory of selection, recombination, and mutation.* Chichester, UK: Wiley.

Burt, C. (1940). *The factors of the mind.* London, UK: University of London Press.

Crow, J. F. (2010). On epistasis: why it is unimportant in polygenic directional selection. *Philosophical Transactions of the Royal Society B*, *365*, 1241–1244.

Crow, J. F., & Kimura, M. (1970). *An introduction to population genetics theory.* New York, NY: Harper and Row.

Daniel, R. M., Kenward, M. G., Cousens, S. N., & De Stavola, B. L. (in press). Using causal diagrams to guide analysis in missing data problems. *Statistical Methods in Medical Research.*

Deary, I. J., Batty, G. D., & Gale, C. R. (2008). Bright children become enlightened adults. *Psychological Science*, *19*, 1–6.

Deutsch, D. (1997). *The fabric of reality: The science of parallel universes and its implications.* London, UK: Penguin.

Dickens, W. T., & Flynn, J. R. (2001). Heritability estimates versus large environmental effects: The IQ paradox resolved. *Psychological Review*, *108*, 356–369.

Do, C. B., Tung, J. Y., Dorfman, E., Kiefer, A. K., Drabant, E. M., Francke, U., et al. (2011). Web-based genome-wide association study identifies two novel loci and a substantial genetic component for Parkinson's disease. *PLoS Genetics*, *7*, e1002141.

Eaves, L. J. (1979). The use of twins in the analysis of assortative mating. *Heredity*, *43*, 399–409.

Eriksson, N., Macpherson, J. M., Tung, J. Y., Hon, L. S., Naughton, B., Saxonov, S., et al. (2010). Web-based, participant-driven studies yield novel genetic associations for common traits. *PLoS Genetics*, *6*, e1000993.

Falconer, D. S. (1985). A note on Fisher's 'average effect' and 'average excess'. *Genetical Research*, *46*, 337–347.

Fisher, R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, *52*, 399–433.

Fisher, R. A. (1941). Average excess and average effect of a gene substitution. *Annals of Eugenics*, *11*, 53–63.

Fisher, R. A. (1952). Statistical methods in genetics. *Heredity*, *6*, 1–12.

Fisher, R. A. (1966). *The design of experiments* (8th ed.). New York, NY: Hafner.

Fisher, R. A. (1970). *Statistical methods for research workers* (14th ed.). New York, NY: Hafner.

Fisher, R. A. (1999). *The genetical theory of natural selection: A complete variorum edition*. Oxford, UK: Oxford University Press.

Freedman, D. A. (1987). As others see us: A case study in path analysis. *Journal of Educational and Behavioral Statistics*, *12*, 101–128.

Freedman, D. A. (2004). Graphical models for causation, and the identification problem. *Evaluation Review*, *28*, 267–293.

Gillies, D. (2001). Review of *Causality*. *British Journal for the Philosophy of Science*, *52*, 613–622.

Glymour, C. (1997). Social statistics and genuine inquiry: Reflections on *The Bell Curve*. In *Intelligence, genes, and success: Scientists respond to* The Bell Curve (pp. 257–280). New York, NY: Springer.

Gould, S. J. (1981). *The mismeasure of man*. New York, NY: Norton.

Greenland, S. (2010). Overthrowing the tyranny of null hypotheses hidden in causal diagrams. In R. Dechter, H. Geffner, & J. Y. Halpern (Eds.), *Heuristics, probability and causality: A tribute to Judea Pearl* (pp. 365–382). London, UK: College Publications.

Hill, W. G., Goddard, M. E., & Visscher, P. M. (2008). Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genetics*, *4*, e1000008.

Hitchcock, C. (2001). Review of *Causality*. *Philosophical Review*, *110*, 639–641.

International Consortium for Blood Pressure Genome-Wide Association Studies. (2011). Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature*, *478*, 103–109.

Jensen, A. R., & Sinha, S. N. (1993). Physical correlates of human intelligence. In P. A.

Vernon (Ed.), *Biological approaches to the study of human intelligence* (pp. 139–242). Norwood, NJ: Ablex.

Johnson, W., Penke, L., & Spinath, F. M. (2011). Heritability in the era of molecular genetics: Some thoughts for understanding genetic influences on behavioural traits. *European Journal of Personality*, *25*, 254–266.

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S.-Y., Freimer, N. B., et al. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, *42*, 348–354.

Kooner, J. S., Saleheen, D., Sim, X., Sehmi, J., Zhang, W., Frossard, P., et al. (2011). Genome-wide association study in individuals of South Asian ancestry identifies six new type 2 diabetes susceptibility loci. *Nature Genetics*, *43*, 984–989.

Laird, N. M., & Lange, C. (2011). *The fundamentals of statistical genetics*. New York, NY: Springer.

Lango Allen, H., Estrada, K., Lettre, G., Berndt, S. I., Weedon, M. W., Fernando, R., et al. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, *467*, 832–838.

Lanktree, M. B., Guo, Y., Murtaza, M., Glessner, J. T., Bailey, S. D., Onland-Moret, N. C., et al. (2011). Meta-analysis of dense genecentric association studies reveals common and uncommon variants associated with height. *American Journal of Human Genetics*, *88*, 6–18.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken, NJ: Wiley.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Lubinski, D., & Dawis, R. V. (1995). *Assessing individual differences in human behavior*. Palo Alto, CA: Consulting Psychologists.

Lunshof, J. E., Bobe, J., Aach, J., Angrist, M., Thakuria, J. V., Vorhaus, D. B., et al. (2010). Personal genomes in progress: From the Human Genome Project to the Personal Genome Project. *Dialogues in Clinical Neuroscience*, *12*, 47–60.

Lynch, M., & Walsh, B. (1998). *Genetics and analysis of quantitative traits*. Sunderland, MA: Sinauer.

MacCorquodale, K., & Meehl, P. E. (1948). On a distinction between hypothetical constructs and intervening variables. *Psychological Review*, *55*, 95–107.

McDonald, R. P. (1985). *Factor analysis and related methods*. Hillsdale, NJ: Erlbaum.

McDonald, R. P. (1986). Describing the elephant: Structure and function in multivariate data. *Psychometrika*, *51*, 513–534.

McDonald, R. P. (1996). Consensus emergens: A matter of interpretation. *Multivariate Behavioral Research*, *31*, 663–672.

McDonald, R. P. (2002). What can we learn from the path equations?: Identifiabiltiy, constraints, equivalence. *Psychometrika*, *67*, 225–249.

McDonald, R. P. (2003). Behavior domains in theory and practice. *Alberta Journal of Educational Research*, *49*, 212–230.

McDonald, R. P. (2010). Structural models and the art of approximation. *Perspectives on Psychological Science*, *5*, 675-686.

McDonald, R. P., & Ho, M. H. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods*, *7*, 64–82.

Meehl, P. E. (1970). Nuisance variables and the ex post facto design. In M. Radner & S. Winokur (Eds.), *Minnesota studies in the philosophy of science vol. IV* (pp. 373–402). Minneapolis, MN: University of Minnesota Press.

Meehl, P. E., & Waller, N. G. (2002). The path analysis controversy: A new statistical approach to strong appraisal of verisimilitude. *Psychological Methods*, *7*, 283–300.

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance.

Psychometrika, *58*, 525–543.

Molenaar, P. C. M. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement: Interdisciplinary Research and Perspective*, *2*, 201–218.

Mulaik, S. A. (2005). Looking back on the indeterminacy controversies in factor analysis. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Contemporary psychometrics: A festschrift for Roderick P. McDonald* (pp. 174–206). Mahwah, NJ: Erlbaum.

Mulaik, S. A. (2010). *Foundations of factor analysis* (2nd ed.). Boca Raton, FL: Chapman and Hall/CRC.

Murray, C. (2002). IQ and income inequality in a sample of sibling pairs from advantaged family backgrounds. *American Economic Review*, *92*, 339–343.

Nisbett, R. E. (2009). *Intelligence and how to get it: Why schools and cultures count.* New York, NY: Norton.

Park, J.-H., Wacholder, S., Gail, M. H., Peters, U., Jacobs, K. B., Chanock, S. J., et al. (2010). Estimation of effect size distributions from genome-wide association studies and implications for future discoveries. *Nature Genetics*, *42*, 570–575.

Patterson, N., Price, A. L., & Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genetics*, *2*, e190.

Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). New York, NY: Cambridge University Press.

Penke, L., Denissen, J. J. A., & Miller, G. F. (2007). The evolutionary genetics of personality. *European Journal of Personality*, *21*, 549–587.

Plomin, R., DeFries, J. C., McClearn, G. E., & McGuffin, P. (2008). *Behavioral genetics* (5th ed.). New York, NY: Worth Publishers.

Price, G. R. (1972). Fisher's 'fundamental theorem' made clear. *Annals of Human Genetics*, *36*, 129–140.

Roberts, B. W., Kuncel, N. R., Shiner, R., Caspi, A., & Goldberg, L. R. (2007). The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. *Perspectives on Psychological Science*, *2*, 313–345.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*, 147–177.

Shipley, B. (2000). A new inferential test for path models based on directed acyclic graphs. *Structural Equation Modeling*, *7*, 206–218.

Shipley, B. (2003). Testing recursive path models with correlated errors using *d*-separation. *Structural Equation Modeling*, *10*, 214-221.

Spearman, C. (1927). *The abilities of man: Their nature and measurement*. New York, NY: Macmillan.

Speliotes, E. K., Willer, C. J., Berndt, S. I., Monda, K. L., Thorleifsson, G., Jackson, A. U., et al. (2010). Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature Genetics*, *42*, 937–948.

Spirtes, P., Glymour, C., & Scheines, R. (2001). *Causation, prediction, and search* (2nd ed.). Cambridge, MA: MIT Press.

Spirtes, P., Richardson, T., Meek, C., Scheines, R., & Glymour, C. (1998). Using path diagrams as a structural equation modelling tool. *Sociological Methods and Research*, *27*, 182–225.

Teslovich, T. M., Musunuru, K., Smith, A. V., Edmondson, A. C., Stylianou, I. M., Koseki, M., et al. (2010). Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*, *466*, 707–713.

Thomson, G. H. (1951). *The factorial analysis of human ability* (5th ed.). London, UK: University of London Press.

Turkheimer, E., & Waldron, M. (2000). Nonshared environment: A theoretical,

methodological, and quantitative review. *Psychological Bulletin*, *126*, 78–108.

Waters, K. M., Stram, D. O., Hassanein, M. T., Le Marchand, L., Wilkens, L. R., Maskarinec, G., et al. (2010). Consistent association of type 2 diabetes risk variants found in Europeans in diverse racial and ethnic groups. *PLoS Genetics*, *6*, e1001078.

Woodward, J. (2003). *Making things happen: A theory of causal explanation.* New York, NY: Oxford University Press.

Wright, S. (1968). *Evolution and the genetics of populations vol. 1: Genetics and biometric foundations.* Chicago, IL: University of Chicago Press.

Author Note

Please send correspondence to jameslee@wjh.harvard.edu. I am particularly grateful to Allan Drummond, Tom Bouchard, and Judea Pearl for their encouragement and generosity.

Footnotes

[1]A causal effect within a given system is *identified* if it can be computed uniquely from any positive probability of the observed variables. Informally, a causal effect is identified if it can be estimated "validly" or "without bias" from the available observations.

[2]Trent Kyono has written a beta version of the program Commentator, which automates many of the analyses demonstrated in this article. Email him at tmkyono@gmail.com.

[3]A basis set consisting of the partial correlations between nonadjacent nodes given their parents always exists if the DAG is *Markovian*—that is, if the only variables connected by bidirectional arcs are *exogenous*, meaning that their causes are unspecified. The model in Figure 4 is Markovian; the only bidirectionally connected nodes have no depicted ancestors. An *endogenous* variable has at least one cause specified in the model; in other words at least one directed edge points into its node. A *Semi-Markovian* model contains at least one bidirectional arc pointing into an endogenous variable, and it may be that a basis set for such a model must contain point predictions that do not take the form of vanishing partial correlations. Critically, it is unknown whether there is a general characterization of a basis set implying all of the point predictions entailed by a semi-Markovian model. Furthermore, when alternative models can be semi-Markovian, entailing the same vanishing partial correlations is only a necessary condition for covariance equivalence to the original model.

Although McDonald (2002) and Shipley (2003) provide methods for semi-Markovian models, these are either tedious to apply or not fully general. This is an area requiring further work. In the meantime the program Commentator does supply *all* point predictions entailed by a semi-Markovian model. For most semi-Markovian DAGs arising in personality research, containing relatively few nodes, a simple and feasible approach to handle the Commentator output is to determine numerically whether a given subset of all

point predictions is in fact a basis set.

Table 1

d-separation tests of the causal model in Figure 4

d-separable nodes	$\hat{\rho}_{ij\cdot\text{parents of } i \cup \text{parents of } j}$ (95% CI)		p-value	
(partial correlation)	Male	Female	Male	Female
parent SES, *antiracism* ($\rho_{16\cdot 234}$)	$-.00$ $(-.036, .030)$	$-.03$ $(-.062, .001)$	.87	.07
SES at age 30, *antiracism* ($\rho_{56\cdot 1234}$)	.06 $(.026, .092)$	$-.03$ $(-.057, .007)$	.0006	.13

Note. $\rho_{ij\cdot\text{parents of } i \cup \text{parents of } j}$ stands for the partial correlation between T_i and T_j given their parents. The p-values in each column can be combined by Fisher's method to provide an overall test of the model for males ($\chi^2_4 = 14.9$, $p < .005$) and females ($\chi^2_4 = 10.1$, $p < .05$).

Figure Captions

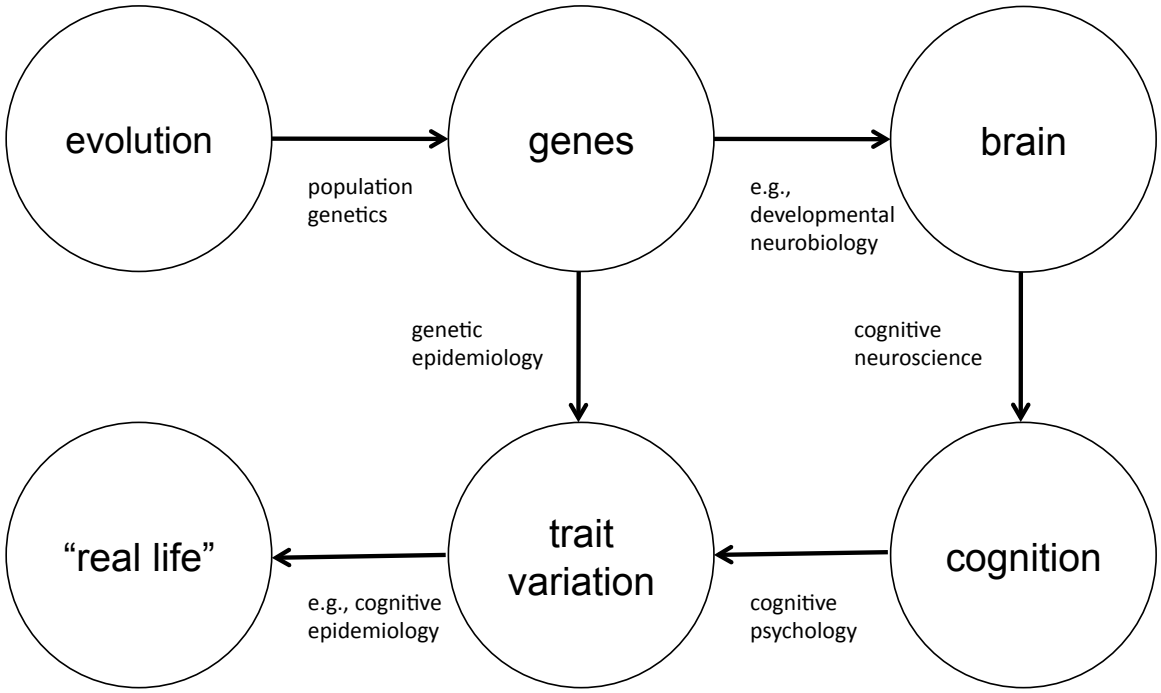
Figure 1. Causal chain hypothesized by some psychologists. This chain happens to be a directed acyclic graph, although it does not represent any formal model. The DAG depicts only some of the possible nodes and edges.

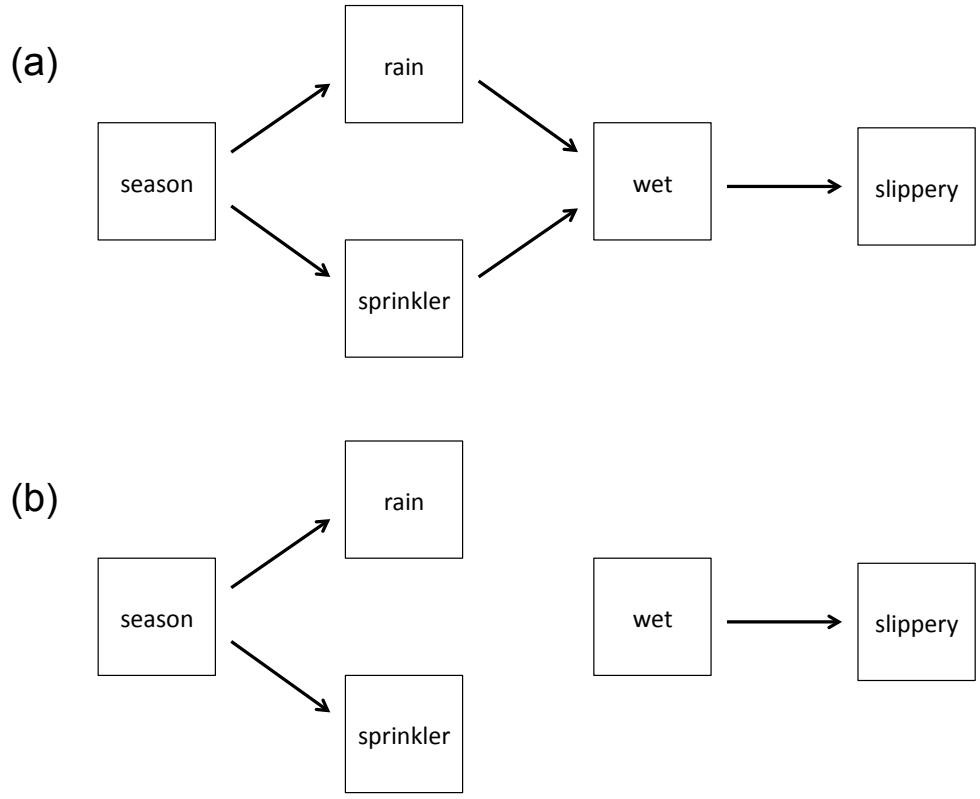
Figure 2. A DAG representing a system (a) before the manipulation of *wet*, and (b) after this manipulation.

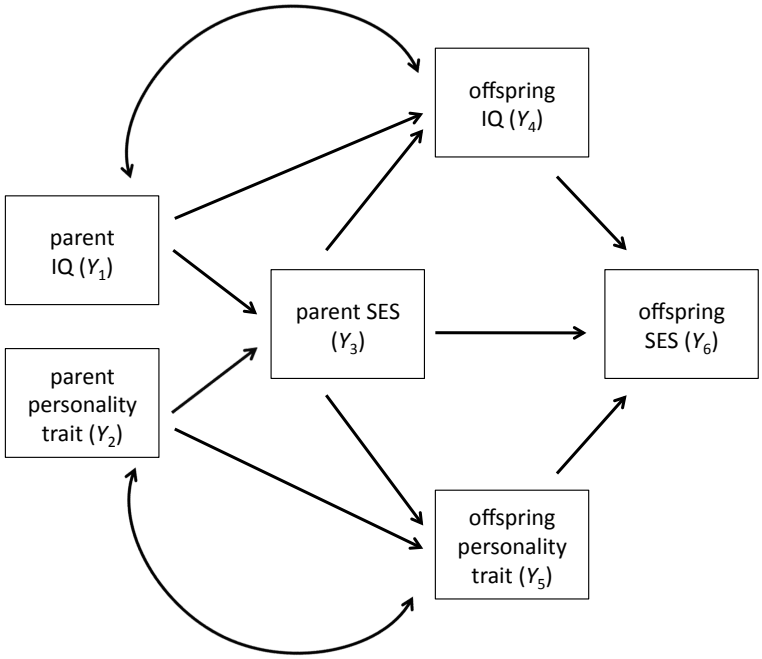
Figure 3. A DAG representing a model of personality and status attainment.

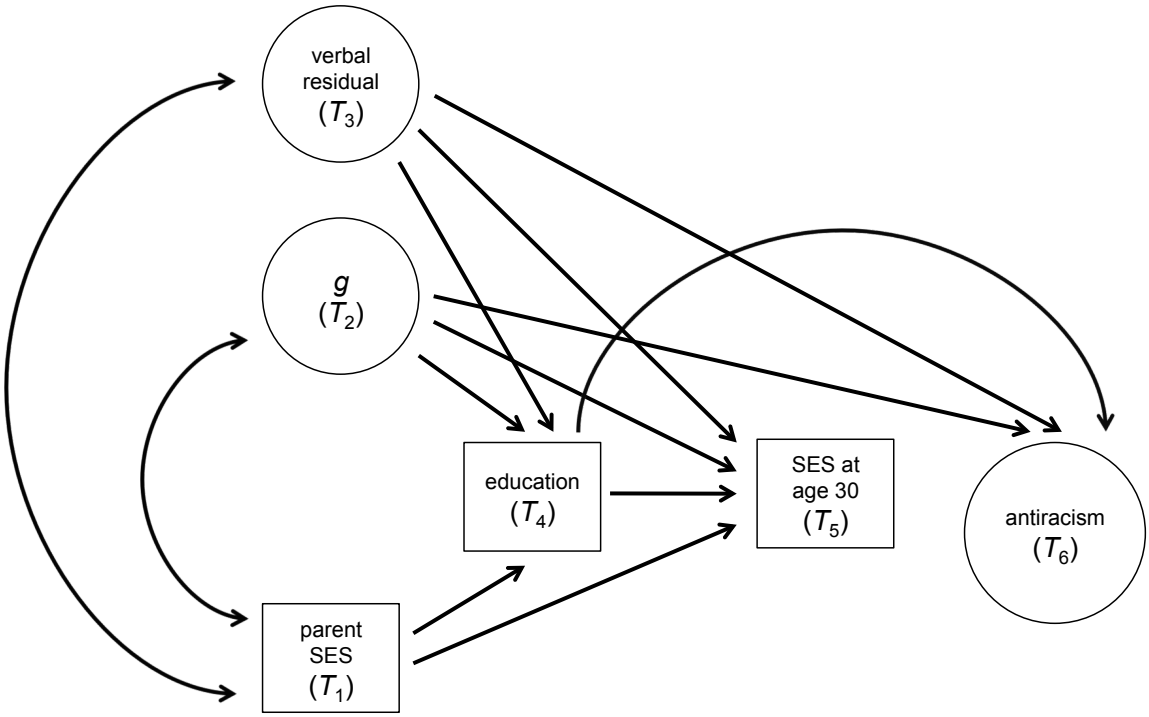
Figure 4. A DAG representing a causal model of the variables studied by Deary et al. (2008).

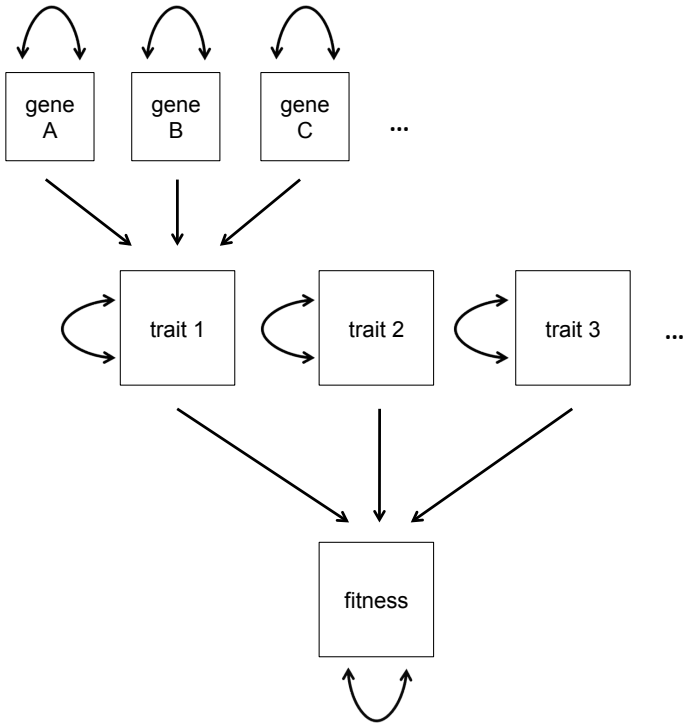
Figure 5. A DAG representing the causal chains from genes to fitness. When considering selection bias, we can simply relabel the bottom node as *appearance in the study.*


evolution
genes
brain
population genetics
e.g., developmental neurobiology
genetic epidemiology
cognitive neuroscience
"real life"
trait variation
cognition
e.g., cognitive epidemiology
cognitive psychology


(a)
season
rain
sprinkler
wet
slippery
(b)
season
rain
sprinkler
wet
slippery


offspring
IQ (Y_4)
parent
IQ (Y_1)
parent SES
(Y_3)
offspring
SES (Y_6)
parent
personality
trait (Y_2)
offspring
personality
trait (Y_5)


verbal residual (T_3)
g (T_2)
education (T_4)
SES at age 30 (T_5)
antiracism (T_6)
parent SES (T_1)


gene A
gene B
gene C
...
trait 1
trait 2
trait 3
...
fitness