Running head: CAUSES AND EFFECTS

Causes and Effects of Common Factors

James J. Lee

Vision Lab, Department of Psychology, Harvard University

Laboratory of Biological Modeling, NIDDK, NIH

Cognitive Genomics Lab, BGI-Shenzhen

Abstract

The comments endorse the usefulness of the graphical framework for causal reasoning in personality psychology. Here I address several recurring themes: (1) details of the graphical framework not explicitly addressed in the target article, (2) the importance of finding a fruitful level of explanation in personality psychology, (3) the problem of selection bias in empirical research, (4) a difference in outlook between nomothetic and idiographic approaches, and (5) whether the causal links between genetic and behavioral variation are indeed empirically tractable.

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

Causes and Effects of Common Factors

I am pleased to find a reasonably firm consensus regarding the utility of the graphical framework for causal reasoning in personality psychology. I divide my response to the comments into five parts, each addressing a recurring theme. At times I express pointed disagreement, but in no way should this be taken as ingratitude toward the praise garnered by my modest contribution.

Further Details of the Graphical Framework

Several comments touch upon further nuances of the graphical framework, which I now take the opportunity to address.

The Back-Door Rule

The target article reproduced a theorem regarding the identification of linear causal effects. This theorem can be regarded as a corollary of what Pearl calls the *back-door theorem*. **Kievit, Waldorp, Kan & Wicherts** cite the back-door theorem in order to emphasize that the hypotheses of the theorem must hold in order for it to be applicable. Unfortunately, their statement of the theorem contains errors and ambiguities. Their hypothesis (i) states that all direct causes of X must be statistically controlled in order to identify the effect of X on Y. What the back-door theorem actually requires, however, is that the set of statistically controlled variables blocks every path between X and Y containing an arrow into X. A member of the blocking set thus need not be a direct cause of X. Moreover, if the path is a colliding path, then the collider—even if it is a direct cause of X—must *not* be a member of the set. **Aspendorf** and I stress that statistically controlling such a variable *opens* its path rather than blocking it.

The necessity of **Kievit's** hypotheses (ii) and (iii) turns on the distinction between

linear and nonlinear models. In the target article, I focused mostly on the linear formulation of important concepts for simplicity. However, since the issue of nonlinearity is not so far in the background of several comments, I will begin expanding on it here.

Kievit's hypothesis (ii), which states that the set of statistically controlled variables cannot include any descendant of X, is actually only a requirement for identifying a *total* effect encompassing all directed paths from X to Y. In a linear system, statistically controlling at least one mediator along an indirect causal path—that is, at least one descendant of X—is necessary to estimate any partial effect.

The expression for the total effect in a nonlinear system is enlightening. If Z d-separates all non-directed paths between X and Y, then the total effect on Y of changing x_1 to x_2 is

$$E\left[(Y \mid do(x_2)) - (Y \mid see(x_1))\right] = \iint y\, p(y \mid x_2, z)\, p(z)\, dz\, dy - \int y\, p(y \mid x_1)\, dy.$$

The form of this expressions tells us that we must average over all possible values of Z; picking just one value may lead to a "stratum-specific" estimand. In other words the identified total effect is an *average* causal effect over the studied population; a treatment that helps one patient may harm another. Also, suppose that we pick different values of X (say x_3 and x_4) in this nonlinear system. Even if $|x_4 - x_3| = |x_2 - x_1|$, the expected change in Y may still depend on the specific values of X.

These observations lead to a succinct characterization of linearity: in a linear causal system, the expected change in Y for a given magnitude of the experimental change in X does not depend on "where we are." That is, the expected change does not depend on the specific values of X, Z, or any other variable. In particular it does not depend on the specific person to whom the manipulation is applied.

In a linear system with disturbances, it should be clear that **Kievit's** hypothesis (iii), which states that the joint probability density is almost everywhere positive, is

indeed satisfied. Even in the nonlinear case, one should drop the "almost" in order to make (iii) a true sufficient condition because a well-chosen set of measure zero with no probability density can prevent the identification of certain *do* quantities.

There are other means besides the back-door rule for determining whether a given set of assumptions identifies a nonlinear causal effect. Pearl has devised a more general calculus for his *do* symbol that allows us to determine whether there is a sequence of transformations eliminating all occurrences of *do* from a given statement. The transformed statement, which contains only instances of *see*, provides an equivalent expression for the desired causal quantity that can be estimated from observational data.

Hypothesis-Free Search Versus Hypothesis Testing

Kievit expresses concern over the number of partial correlations (conditional independencies) that must be tested in applications. My *g*-antiracism example, however, required the testing of only two partial correlations. Whereas that example demonstrated the testing of an *a priori* causal model, **Kievit** must have in mind the search for a model based on nothing more than an algorithmic evaluation of all possible partial correlations. An entire research program is dedicated to this kind of hypothesis-free approach (Spirtes, Glymour, Scheines, & Tillman, 2010). My impression is that causal search algorithms do not work well when applied to real data, possibly in part because of the multiple-testing issue that **Kievit** mentions. Whenever analysts have prior knowledge that would be privileged over contradictory results delivered by search algorithms, it seems that they should simply use it. Readers are encouraged to explore this issue for themselves.

Although inevitably sounding pedantic, I must point out that **Kievit's** combinatorial calculation for a five-node graph is erroneous. $5!/(3!\,2!)$ is equal to ten, not twenty. And for any given pair of variables, the number of possible partial correlations (including the zero-order correlation) is eight, not six.

Foundational Issues

Steyer poses two specific challenges to my claim that Pearl's graphical theories (or their mathematical equivalents) present the correct formal framework for what we mean by the word *causality*. A general attitude along these lines is evident in several other comments.

First, **Steyer** attacks the notion of a causal effect of X on Y as the change that Y would undergo if X were randomized, expressing dissatisfaction with its apparent inapplicability in cases where X is not (or cannot) be experimentally controlled. This objection, however, confuses definitions and empirical operations. Imagining the thought experiments implied by each directed edge in a DAG can sharpen our justifications for including certain arcs in a conjectural model and deciding what kind of arcs they should be. But this does not imply that sensitivity to actual or potential human manipulation *defines* causation.

> Surely we have causation without manipulation. The moon causes tides, race causes discrimination, and sex causes the secretion of certain hormones and not others. Nature is a society of mechanisms that relentlessly sense the values of some variables and determine the values of others; it does not wait for a human manipulator before activating those mechanisms. (Pearl, 2009, p. 361)

I believe that **Steyer's** more specific point regarding indirect or conditional causal effects fails for the same reason. In a randomized experiment examining the effect of sleep deprivation on neuroticism, say, the average causal effect of the treatment may be identified whereas more specific causal effects may not. For example, it may not be possible to identify how the treatment affected a particular person (**Kievit**; **Borsboom, van der Sluis, Noordhof, Wichers, Geschwind, Aggen, Kendler & Cramer**). But that does not mean that the person was not affected somehow. A *do* operation on the

appropriate DAG mathematically defines this effect regardless of whether the DAG's structure permits particular "real-life studies" to know what this effect is. In fact, much DAG theory is devoted to precisely these concerns (**Pearl**).

Second, **Steyer** joins **Jackson and Spain** in accusing the graphical framework of conflating the disturbance term in a causal equation with the error term in a least-squares regression. It is a geometric fact of least squares that the error term must be uncorrelated with the predictor. Because it cannot be true in general that the measured causes of a certain effect are independent of its unmeasured causes, **Steyer** believes that he has undermined Pearl's arguments for the depth and essentiality of his approach. But on the very page from which **Steyer** quotes, Pearl clearly allows a semi-Markovian DAG to include correlated disturbances. If this were not the case, the various theorems giving the circumstances under which a causal equation *is* a regression would often have a rather trivial flavor. This oversight may be responsible for the puzzling remarks with which **Steyer** follows his quotation.

Steyer (1984) claims to offer an alternative account of causality based on probability theory. I unfortunately found this account to be quite obscure. For example, the condition given there for the identification of a causal effect strikes me as the very definition of a conditional expectation. Instead of dwelling on details, I will offer some reactions based on first principles. Pearl has convinced me that probability theory is *inherently* incapable of representing causal notions. Why might this be? One of my statistics instructors once defined his subject as the use of finite data to infer the parameters of probability distributions, and indeed the first volume of the *Advanced Theory of Statistics* is dedicated exclusively to the properties of probability distributions (Stuart & Ord, 1987). Now suppose that the rows of a given data matrix go to infinity, allowing us to estimate the parameters of the relevant probability distribution (and functions thereof) as precisely as we please. We thus have in our hands error-free estimates of means, variances, higher

moments, correlations, odds ratios, principal components, propensity scores, Granger-"causality" coefficients, and so on. Have we gone any way toward understanding the causal mechanisms generating the data? One could fairly reply: *not yet*. To say something about the causal process inducing the obtained distribution, we must invoke assumptions about matters that are inherently non-statistical: *randomization*, *confounding*, *selection*, and so forth.

> Take the concept of randomization—why is it not statistical? Assume we are given a bivariate density function $f(x, y)$, and we are told that one of the variables is randomized; can we tell which one it is by just examining $f(x, y)$? Of course not; therefore, following our definition, randomization is a causal, not a statistical concept. . . .
>
> Note, however, that the purpose of the causal–statistical demarcation line . . . is not to exclude causal concepts from the province of statistical analysis but, rather, to encourage investigators to treat causal concepts distinctly, with the proper set of mathematical and inferential tools. Indeed, statisticians were the first to conceive of randomized experiments, and have used them successfully since the time of Fisher (1926). However, both the assumptions and conclusions in those studies were kept implicit, in the mind of ingenious investigators; they did not make their way into the mathematics. For example, one would be extremely hard pressed to find a statistics textbook, even at the graduate level, containing a mathematical proof that randomization indeed produces unbiased estimates of [causal] quantities . . . (Pearl, 2009, p. 332)

These considerations leave me skeptical of attempts to build a formal account of causality on purely probabilistic grounds.

The foundational importance and distinctiveness of causal notions implies that the

graphical framework (or a mathematical equivalent) is *always* employed, at least implicitly, whenever causal inferences are drawn. **Jackson and Spain** present instruments and propensity scores as alternatives to the graphical approach, but in fact the graphical approach subsumes both of these concepts. Although I am admittedly not a methodologist, Pearl's (2009) treatments of instruments and propensity scores may be the most lucid that I have seen anywhere in the literature.

I actually share the reservations of **Kievit**, **Borsboom**, and **Jackson and Spain** regarding cross-sectional studies of high-level variables that do not incorporate some special feature such as those discussed in the target article (families, genetics, natural randomization). Longitudinal tracking of individuals should be added to this repertoire. In fact, one reason why the target article did not treat cyclic models is that I join Shipley (2000) in suspecting that a cyclic model can usually be reduced to an acyclic or block-acyclic model by sufficiently fine-grained distinctions among time points within individuals. But the importance of causal notions does not diminish when we turn our attention from a large cross-section of a population at a single time point to a small number of individuals across many small points. Note that each directed edge in Cramer et al.'s (this issue) Figure 5 represents temporal order rather than a cause-effect relation. In order to make the leap from these lagged correlations to cause and effect, we need to invoke randomization, instruments, confounding, selection bias, or other members of the conceptual family surrounding *d*-separation.

I take issue with **Borsboom's** claim that empirical application of graphical theories "has not been very successful to date." This misconception may be based on artificially restricting the content of the graphical framework to narrow applications such as causal search algorithms.

It should be reemphasized that possessing a clear and formal notion of causation does not necessarily enable us to discover the causal structure of any particular system.

The situation is analogous to the failure of formal logic to resolve certain philosophical disputes (Gensler, 2002). If the truth of the assumptions or the meanings of key terms are in doubt, then we cannot draw any firm conclusions. But such failures are no reason to denigrate *logic*. The logic of causality is no different in this respect. Similarly, if *any* deductions are warranted at all, then their warrant must rest on both the requisite assumptions and the appropriate governing logic. Causal deductions are again no different.

For instance, how do we know that smoking kills? Let us examine one piece of the "nomological net" (**Jackson & Spain**). Strong evidence against Fisher's hypothesis that certain genotypes cause both disease and a personality disposed to smoking comes from studies of smoking-discordant monozygotic twins in which the smoker tended to die first (Kaprio & Koskenvuo, 1989; Carmelli & Page, 1996). Now why is this strong evidence? Letting G stand for genotype, we have

$$P(\mathit{death},\ \mathit{smoking} \mid G = g) = P(\mathit{death},\ \mathit{smoking} \mid \mathit{see}(G = g)),$$

but wish to infer

$$P(\mathit{death},\ \mathit{smoking} \mid \mathit{do}(G = g)).$$

This substitution is legitimate because it was nature that *fixed* G to be the same for both members of a twinship. That is, because the same physical process produced the genotype of each twin, the nodes in the causally prior subgraph have no variation that can be transmitted to smoking and death; therefore, by Rule 2 of Pearl's *do* calculus, we can safely delete all directed edges converging on G. This is quite unlike matching by propensity scores, say, where individuals in the same stratum are merely *observed* to have the same value of the propensity score (a function of the matching variables). Further assumptions regarding the causal processes outputting the matching variables, such as the applicability of the back-door rule, must be justified before the *see* in the latter case can be replaced with *do*. Even if all this is already intuitively clear to some, we can only profit

from the explication and systematization of *ad hoc* intuition.

How do **Borsboom** and **Johnson** explain the success of genome-wide association studies (GWAS)? Given the replication of GWAS results across nations and racial groups—and, most importantly, within families—it has become clear that the "batting average" of causal inference with this technique is well above .500. This success rate should make GWAS the envy of other biomedical, behavioral, and social scientists who must attempt to advance causal claims on the basis of observational data. We are thus entitled to ask: what special features of GWAS make causal inference, if not infallible, at least reliable? My own attempt to provide an answer is of course a *post hoc* explanation rather than an *a priori* justification by the investigators themselves, but the timestamp on the argument is irrelevant to the basic principle that specifying the DAG containing the putative cause and effect, convincing others of its validity, and demonstrating the identification of the desired quantity is essential to the justification of any empirical study claiming to advance our causal knowledge.

> The sharp distinction between statistical and causal concepts can be translated into a useful principle: behind every causal claim there must lie some causal assumption that is not discernable from the joint distribution and, hence, not testable in observational studies. . . .
>
> Any causal premise that is cast in standard probability expressions, void of graphs, counterfactual subscripts [in the language of the Neyman-Rubin potential-outcome framework], or $do(*)$ operators, can safely be discarded as inadequate.
>
> While this harsh verdict may condemn valuable articles in the empirical literature to the province of inadequacy, it can save investigators endless hours of confusion and argumentation . . . More importantly, the verdict should

> encourage investigators to visibly explicate causal premises, so that they can be communicated unambiguously to other investigators and invite professional scrutiny, deliberation, and refinement. (Pearl, 2009, p. 40, 334)

This is why it is somewhat misguided to dwell on "the difficulty of explicating all assumptions about which variables are measured, their causal relations, and possibility of co-occurring events when applying graphical modeling" (**Kievit**), as if there were a non-graphical alternative that avoided these problems. As Pearl (2009, pp. 173–200) shows in his *tour de force* treatment of Simpson's paradox, alternative frameworks do not provide any insight into issues such as sign reversals across levels of aggregation and frequently aggravate confusion. Regrettably, **Kievit's** own comment bears this out; his example of juggling practice and brain volume, despite being brief, contains a number of ambiguities. I anticipate that basic graphical notions (such as the distinction between *see* and *do*) will strike future scientists as no more distinctively "Pearlian" than basic statistical notions (such as the distinction between an estimate and a parameter) strike us nowadays as distinctively "Fisherian." These ideas will have become so ingrained that to attribute them to Pearl will seem akin to attributing "the invention of the wheel to Mr. So-and-So" (Savage, 1976, p. 450).

Common Factors and Levels of Analysis

> • Justice Scalia has the intellect to be a leader of the United States Supreme Court's conservative wing, but his surly temperament has prevented him from assuming this mantle.

The graphical framework now provides tools to reason precisely about previously difficult concepts such as *necessity*, *sufficiency*, and *prevention* that are inherent in this statement (**Pearl**). Remarkably it now seems that the invocation of attributes such as *intellect*, *conservatism*, and *surliness* is the more controversial issue. Even as I maintain that

common factors are an attempt to quantify folk-psychological attributes of this kind, others continue to question the usefulness of this enterprise (**Borsboom**).

I will now present a highly idealized analogy to *sprinting ability* in the hope of clarifying the points on which the commentators and I appear to agree. It seems reasonable to believe that a single measurement cannot establish the rank order of any population in sprinting ability; we conceive of this ability as an average over a certain class of performances. The boundaries of this class are uncertain. Is the 800 meters a sprint? Nevertheless there are certain performances that clearly belong to the core of the class. Now we can think of a population's rank order in "sprinting scores" as the limiting rank order approached as we aggregate replications of performances within the class. If the correlations among different sprints are very high, then, as is often said in applied mathematics, infinity might not be far away. But the near determinacy of sprinting scores does not imply that there is any single physical attribute of an individual's body in correspondence with his score. There *might* be such an attribute. If it turned out that number of fast-twitch muscle fibers was perfectly correlated with sprinting scores, then we might drop the concept of sprinting ability and refer instead to muscle fibers. But such an identity is extremely unlikely. It is far more plausible that sprinting ability is an abstraction emerging from interactions among the skeletal, muscular, circulatory, and respiratory systems. Sprinting ability would not even appear as a node in a DAG specifying the low-level causes of 100-meter performance in complete detail.

Is sprinting ability nevertheless a useful concept? It certainly seems so. We can say, for example, that some positions in football require more sprinting ability than others. We can talk about genetic variants affecting sprinting ability, and in fact there *are* such variants (MacArthur et al., 2007). It therefore seems that at least certain high-level traits can be scientifically useful, either as placeholders in the course of reductionistic research or as fundamental entities in their own right. This was a point that I tried to convey with

my long quotation of the theoretical physicist David Deutsch. The question that seems to trouble the commentators, then, is whether common factors (imperfectly measured folk-psychological traits) are useful high-level "cartographic" features to have singled out from the teeming landscape of individual differences (**Condon, Brown-Ridell, Wilt & Revelle**). Meehl (1978) admittedly regarded this task of parsing "continuous streams of observable behavior into chunks that can be reasonably well measured and organized" as an open problem rather than a *fait accompli* of psychological science.

Condon believes that the target article provides no "metric" for evaluating proposed solutions to this problem. It seems to me that no satisfactory metric of this kind has been proposed in any scientific discipline where the problem of "murdering to dissect" has arisen (Wagner, 2001; Blows, 2007), and it may indeed be in the nature of the problem to be without a generally applicable solution. For now the evaluation of "construct validity" remains one of the case-by-case, non-automatable aspects of behavioral science to which **Weiss and Bates** refer. I will point out, however, that the successful embedding of a common factor in a causal model has long been thought to support its validity (Cronbach & Meehl, 1955; Messick, 1989). Now it is quite plausible that the low-level causes of behavioral variation are arranged in a complex cyclic graph (van der Maas et al., 2006). But can we reasonably approximate this situation with a block-acyclic graph, containing d-separable nodes, in which the common factor plays the role of a cause or effect? If so, this suggests that we have indeed carved out a useful high-level attribute of the respondents. There is a suggestion of the relation between statistical mechanics and thermodynamics, although this analogy should not be carried too far.

I will add one more thought to emphasize the likely idiosyncratic nature of trait validation. By adding up randomly chosen questionnaire items and anthropometric measurements, one can put together a wholly arbitrary and trivial phenotype. In fact, many off-hand criticisms of IQ tests have leveled this charge of aggregating an atheoretical

hodgepodge. Since such a Borgesian construct will be heritable if any of its summands are heritable (Johnson, Penke, & Spinath, 2011), a sufficiently well-powered GWAS will certainly find genetic variants affecting it. The upshot is that successful GWAS results are not enough to ensure that we have measured a useful trait. But what if such results indicate a disproportionate clustering of causal variants in particular biological pathways (Lee et al., 2012)? Furthermore, given recent advances in the sequencing of ancient DNA (Rasmussen et al., 2010; Green et al., 2010; Reich et al., 2010; Keller et al., 2012) and the prediction of additive genetic values (Goddard, 2009; Yang, Lee, Goddard, & Visscher, 2011), we may conceivably be able to estimate the level of g that an ancient hominin would have obtained if reared in a modern society. What would we make of a finding that a Neandertal individual had a level of g three standard units below the current mean?

Although this thought experiment regarding biological significance does not seem to illustrate any algorithmic principle of validation, it suggests that uses to common factors may not necessarily be limited to verifications of what many regard to be common sense (perhaps mistakenly). Followers of American politics need no scientific formality to believe that the statement regarding Justice Scalia is probably true. Some might say that intelligence tending to promote liberalism is also obvious; left-liberals and libertarians alike often claim that their views are compelled by reason. A glib naysayer might dismiss much of the literature on construct validity as demonstrating correlations with external variables that common sense already tells us are correlated with the attribute that the scale is intended to measure. The kinds of biological insights that might follow from genetic research, however, are certainly no matters of common sense.

The identification of common factors with psychological attributes arguably depends on the assumption of an infinite domain of items measuring a "dominant" or "essentially single" dimension (Stout, 1990; McDonald, 1999). This is admittedly a very rarified idealization, and it is perhaps understandable if **Borsboom** declines to adopt it. My own

judgment, which is also informed by unpublished data, is that the idealization holds tolerably well for cognitive abilities (Lawrence & Dorans, 1987; Cook, Dorans, & Eignor, 1988; Humphreys, Lubinski, & Yao, 1993; Segall, 2001).

The Problem of Selection Bias

I agree with **Aspendorf** that selection bias is a taboo subject among behavioral scientists. It was selection bias, more than any other topic, that provoked colleagues who read an early draft of the target article to criticize it as "negative," "destructive," and "pessimistic." An important aim of the target article is to bring the problem of selection bias into daylight.

Insert Figure 1 about here

Jackson and Spain do not distinguish selection bias from confounding. Figure 1 hopefully clarifies the difference. Refraining from any assumptions in an attempt to discover whether X causes Y, we must allow unknown confounders to affect these variables and selection into the study to be affected by them in turn. Assuming that a specific random mechanism affects X (and X alone) would suffice to identify any $X \rightarrow Y$ causal effect in the absence of selection bias. However, given the presence of selection bias, more assumptions are necessary regarding the d-separation of all paths between X and Y passing through study appearance. Bareinboim and Pearl (in press) study this issue in detail. One *sufficient* assumption is that whether a participant appears in the study is also determined by a random mechanism. A random sample of individuals who are currently accessible will satisfy this assumption in some cases. However, if death, disability, or the like have removed certain individuals from the pool of potential participants, then even random sampling from the pool may not be enough. This difficulty

admittedly complicates studies of aging.

Figure 1 suggests a succinct formulation of confounding and selection bias: confounding is any contribution to the association between putative cause (X) and effect (Y) that can be removed by randomization of X, whereas selection bias is any contribution that can be removed by randomization of appearance in the study.

I join **Aspendorf** in calling for empirical reports to discuss whether selection bias might unduly affect the results. We seem to be full agreement that the graphical framework's highlighting of selection bias should not be construed as a negative contribution; any and all insights into the roadblocks on the way to causal knowledge are welcome. I will paraphrase John F. Kennedy: "We do not do science because it is easy; we do it because it is hard."

Nomothetic and Idiographic Orientations

Kievit and **Borsboom** take me to task for failing to distinguish between within-person causal effects and across-person correlations. Although the graphical framework does in fact treat the issues of transportability across environments (countries differing in climate and infrastructure) and nonlinearity (causal effects depending on "where we are") in complete generality, perhaps the target article should have discussed these matters in more detail, particularly because it is not always obvious how **Pearl's** mathematical treatments connect to statistical issues in finite samples. It seems to me, however, that a more fundamental issue divides us.

Suppose that we could manipulate Antonin Scalia's level of g at the age of 11 so as to place him even further in the right tail of his cohort's ability distribution. (Recalling the example of sprinting ability, we can be fairly confident that there is no single physical "thing in the head" to be altered in order to bring about this change.) We then wait until he has reached 30 years of age before asking him to fill out a questionnaire regarding his

social and political views. Will this counterfactual Scalia be less conservative than the actual Scalia? Perhaps. But it also seems reasonable to suppose that the counterfactual Scalia might be *more* conservative. His increased g may allow him to become more adept at rationalizing positions that he holds for nonintellectual reasons and attacking the arguments of his liberal opponents. **Kievit** and **Borsboom** express an intense interest in these kinds of idiosyncratic dynamics; they want to know why a *particular* person is the way that he is.

I concede that linear models, which aim to approximate average causal effects (see my earlier discussion of the back-door rule), are idealized simplifications that may distort the dynamics of an individual personality over the lifespan. One argument for the primacy of an idiographic orientation is that an average causal effect in a given population seems to lack the invariance property one might expect from a useful lawlike generalization. As the composition and environment of the population changes, perhaps under the influence of the very causal system under consideration, the average causal effect will change and conceivably even switch sign.

But there is a compelling argument for the other side. Ronald Fisher was well aware that the same allelic substitution may bring about different phenotypic effects in different individuals. This might happen, for example, because of differences between two people at other loci affecting the trait (**Johnson**). Nevertheless Fisher thought that the best linear predictions of genotypic values were more fundamental than actual genotypic values—so much so that he called the former *true* genetic values and conceived of the discrepancies as substantively unimportant errors (Fisher, 1918, 1999). In other words Fisher was already retreating from the ideal of deterministic ("mechanistic") prediction mentioned by **Condon**. It is not clear that anyone has fully understood Fisher's reasoning on these matters, but one important motivation for his view seems to have been the fact that the number of possible genotypes greatly exceeds the number of allele frequencies on which

they depend (Fisher, 1941). For L causal loci there are 3^L multilocus genotypes. If L is equal to 23—one trait-affecting locus per chromosome—the number of possible genotypes already exceeds the current size of the human population by a factor of 10. It is now apparent that traits such as g and schizophrenia are affected by thousands of loci, which means that the number of possible genotypes dwarfs the number of protons in the observable universe. These calculations harbor some surprising consequences. For instance, a genotype that is relatively probable, in the sense that its constituent alleles are common in the population, will often fail to have a single exemplar.

Given the enormous difficulty facing any attempt to elucidate the complete nonlinear genotype-phenotype mapping, the nomothetic bias inherent in the theoretical importance that Fisher attached to the average effect of an allelic substitution poses some obvious advantages in practicality and economy of thought. Moreover, whether the average effects suffer from a lack of invariance is not an all-or-nothing matter; both the relative magnitude of the additive genetic variance and the transportability of genetic findings across populations provide checks on the degree to which the average effects tend to be "slowly varying functions" of the causal background. Thus, in the context of genetic research at least, a nomothetic orientation has a sound rationale backed by an impressive and mounting record of empirical success (Waters et al., 2010; Lango Allen et al., 2010; Yang, Manolio, et al., 2011; Goddard, Lee, Yang, Wray, & Visscher, 2011; Davies et al., 2011; International Consortium for Blood Pressure Genome-Wide Association Studies, 2011; Kooner et al., 2011; Lanktree et al., 2011; Lee et al., 2012).

Kievit and **Borsboom** contemplate longitudinal studies where the putative causes, unlike genotypes, show variation within individuals across time. It is not clear to me that this approach avoids any problems of combinatorial explosion or high dimensionality. Even if we have enough replications over time to establish that a certain person consistently avoids joining stampedes out of the allegedly burning theaters, the question

remains as to *why* some people flee and others freeze. By raising these issues, however, I do not mean to imply that the mere presence of difficulties should deter this ambitious and worthwhile research program. In recent years we have perhaps neglected Allport's (1937) vision of an all-encompassing personality psychology, unduly emphasizing the nomothetic over the idiographic. We should not lose contact with either approach.

Causal Inference in Gene-Trait Association Studies

I regrettably do not understand **Johnson's** argument. Does she claim that most GWAS results from studies of unrelated individuals are false positives? Does she deny that family studies whose causal assumptions invoke only Mendel's laws are immune to confounding? Her reference to gene expression is puzzling because detailed studies of gene expression are often used to trace the intermediate mechanisms between genetic variants identified in mapping studies and the focal phenotype (Pomerantz et al., 2009; Musunuru et al., 2010; Visser, Kayser, & Palstra, 2012). I believe that **Johnson** unintentionally reveals the disciplinary value in supplementing verbal arguments with graphical ones. Naturally I disagree that the transparency of the graphical framework "lies completely in the hands of the researcher." Whereas statistical assumptions are often poorly understood and thus insufficiently scrutinized, the DAG framework allows causal assumptions to be conveniently visualized and therefore readily challenged by critics. The ease with which DAGs stimulate the proclivity of scientists to evoke alternative causal scenarios is one of the graphical framework's attractive features.

With the phrase "ultimate causality," **Johnson** seems to mock the target article's emphasis on genetics. She believes that the fallacy of pointing to genetics—of all things—as a clean system for the isolation of cause and effect is so obvious that she sees no need to accompany it with any detailed argument or even a bare mention of specific genetic phenomena.

There seems to be little point in rehashing the target article's arguments in the face of such vague innuendos. Instead I will reiterate my deepening conviction that there is indeed a special connection between genetics and the notion of causality. Long before Darwin, biologists had already marveled at the exquisite adaptation of organisms to their natural environments. To capture what we mean by adaptation a little more precisely, we can conceive of an organism's actual mean phenotype as a point in a high-dimensional space and the optimal phenotype given the organism's environment as another point in this same space (Fisher, 1999). In this conception adaptation is a high correlation between the coordinates of these two points. Now recall the taxonomy of correlations, which states that a non-coincidental correlation between X (phenotype) and Y (environment) must be attributable to

I. X causing Y,

II. Y causing X,

III. X and Y being effects of a common cause, or

IV. X and Y being causes of a common effect that has been statistically controlled.

We can determine which of these four explanations encompasses the phenomenon of biological adaptation. Pre-Darwinians often invoked explanation III, ascribing the role of common cause to a benevolent Creator. After Darwin and Mendel, some biologists continued to believe in explanation II, invoking hypothetical mechanisms by which environmental circumstances might mold the causes of phenotypic variation. Various theories of Lamarckianism and directed mutation fall under this type. Weismann (1893) gave reasons for rejecting this class of explanations that remain cogent today. Explanation I, in which organisms seek out or create environments promoting their own fitness, obviously cannot suffice because such niche-seeking capacities are themselves complex

adaptations. It was the genius of Darwin to realize the power of explanation IV: phenotypes and environments cohere in such an uncanny way because nature is a statistician who has allowed only a subset of the logically possible combinations to persist over time.

Although phenotypes are what nature selects, it cannot be phenotypes alone that preserve the record of natural selection. Phenotypes typically lack the property that variations in them are replicated with high fidelity across an indefinite number of generations. DNA, however, *does* have this property—-hence the memorable phrase "the immortal replicator" (Dawkins, 1976). If DNA is furthermore causally efficacious, such that the possession of one variant rather than another has consequences for the state of the world that are reasonably robust, then we have the potential for natural selection to bring about a lasting correlation between environmental demands and the causes of adaptation to those very same demands.

When statistically controlling fitness, nature does not actually use the average *effect* of any allele. If an allele has a positive average *excess* in fitness, for any reason whatsoever, it will tend to displace its alternatives. Nevertheless it seems to be the case that nature correctly picks out alleles for their effects often enough; the results are evident in the living world all around us. I am confident that where nature has succeeded, patient and ingenious human scientists will be able to follow.

References

Allport, G. W. (1937). *Personality: A psychological interpretation*. New York: Holt.

Bareinboim, E., & Pearl, J. (in press). Controlling selection bias in causal inference. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics*.

Blows, M. W. (2007). A tale of two matrices: Multivariate approaches in evolutionary biology (with discussion). *Journal of Evolutionary Biology*, *20*, 1–44.

Carmelli, D., & Page, W. F. (1996). Twenty-four year mortality in World War II US male veteran twins discordant for cigarette smoking. *International Journal of Epidemiology*, *25*, 554–559.

Cook, L. L., Dorans, N. J., & Eignor, D. R. (1988). An assessment of the dimensionality of three SAT-Verbal test editions. *Journal of Educational Statistics*, *13*, 19–43.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281–302.

Davies, G., Tenesa, A., Payton, A., Yang, J., Harris, S. E., Liewald, D., et al. (2011). Genome-wide association studies establish that human intelligence is highly heritable and polygenic. *Molecular Psychiatry*.

Dawkins, R. (1976). *The selfish gene*. New York: Oxford University Press.

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

Gensler, H. J. (2002). *Introduction to logic*. London: Routledge.

Goddard, M. E. (2009). Genomic selection: Prediction of accuracy and maximisation of

long term response. *Genetica*, *136*, 245–257.

Goddard, M. E., Lee, S. H., Yang, J., Wray, N. R., & Visscher, P. M. (2011). Response to Browning and Browning. *American Journal of Human Genetics*, *89*, 193–195.

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., et al. (2010). A draft sequence of the Neandertal genome. *Science*, *5979*, 710–722.

Humphreys, L. G., Lubinski, D., & Yao, G. (1993). Utility of predicting group membership and the role of spatial visualization in becoming an engineer, physical scientist, or artist. *Journal of Applied Psychology*, *78*, 250–261.

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.

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.

Kaprio, J., & Koskenvuo, M. (1989). Twins, smoking and mortality: A 12-year prospective study of smoking-discordant twin pairs. *Social Science and Medicine*, *29*, 1083–1089.

Keller, A., Graefen, A., Ball, M., Matzas, M., Boisgeurin, V., Maixner, F., et al. (2012). New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing. *Nature Communications*, *3*, 698.

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.

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.

Lawrence, I. M., & Dorans, N. J. (1987). *An assessment of the dimensionality of the SAT-Mathematical.* Presented at the Annual Meeting of the National Council on Measurement in Education.

Lee, S. H., DeCandia, T. R., Ripke, S., Yang, J., Schizophrenia Psychiatric Genome-Wide Association Study Consortium, International Schizophrenia Consortium, et al. (2012). Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nature Genetics*, *44*, 247–250.

MacArthur, D. G., Seto, J. T., Raftery, J. M., Quinlan, K. G., Huttley, G. A., Hook, J. W., et al. (2007). Loss of *ACTN3* gene function alters mouse muscle metabolism and shows evidence of positive selection in humans. *Nature Genetics*, *39*, 1261–1265.

McDonald, R. P. (1999). *Test theory: A unified treatment.* Mahwah, NJ: Erlbaum.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*, 806–834.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13–103). New York: Macmillan.

Musunuru, K., Strong, A., Frank-Kamenetsky, M., Lee, N. E., Ahfeldt, T., Sachs, K. V., et al. (2010). From noncoding variant to phenotype via *SORT1* at the 1p13 cholesterol locus. *Nature*, *466*, 714–719.

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

Pomerantz, M. M., Ahmadiyeh, N., Jia, L., Herman, P., Verzi, M. P., Doddapaneni, H., et al. (2009). The 8q24 cancer risk variant rs6983267 shows long-range interaction with

MYC in colorectal cancer. *Nature Genetics*, *41*, 882–884.

Rasmussen, M., Li, Y., Lindgreen, S., Pedersen, J. S., Albrechtsen, A., Moltke, I., et al. (2010). Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature*, *463*, 709–840.

Reich, D., Green, R. E., Kircher, M., Krause, J., Patterson, N., Durand, E. Y., et al. (2010). Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature*, *468*, 1053–1060.

Savage, L. J. (1976). On rereading R. A. Fisher (with discussion). *Annals of Statistics*, *4*, 441–500.

Segall, D. O. (2001). General ability measurement: An application of multidimensional item response theory. *Psychometrika*, *66*, 79–97.

Shipley, B. (2000). *Cause and correlation in biology: A user's guide to path analysis, structural equations and causal inference*. Cambridge, UK: Cambridge University Press.

Spirtes, P., Glymour, C., Scheines, R., & Tillman, R. (2010). Automated search for causal relations: Theory and practice. In R. Dechter, H. Geffner, & J. Y. Halpern (Eds.), *Heuristics, probability and causality: A tribute to Judea Pearl* (pp. 467–506). London: College Publications.

Steyer, R. (1984). Causal linear stochastic dependencies: The formal theory. In E. Degreef & J. van Buggenhaut (Eds.), *Trends in mathematical psychology* (pp. 317–346). Amsterdam: Elsevier.

Stout, W. F. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika*, *55*, 293–325.

Stuart, A., & Ord, J. K. (1987). *Kendall's advanced theory of statistics vol. 1: Distribution theory* (5th ed.). New York: Oxford University Press.

van der Maas, H. L. J., Dolan, C. V., Grasman, R. P. P. P., Wicherts, J. M., Huizenga,

H. M., & Raijmakers, M. E. J. (2006). A dynamical model of general intelligence: The positive manifold of intelligence by mutualism. *Psychological Review*, *113*, 842–861.

Visser, M., Kayser, M., & Palstra, R.-J. (2012). *HERC2* rs12913832 modulates human pigmentation by attenuating chromatin-loop formation between a long-range enhancer and the *OCA2* promoter. *Genome Research*, *22*, 446–455.

Wagner, G. P. (2001). *The character concept in evolutionary biology.* San Diego, CA: Academic Press.

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.

Weismann, A. (1893). The all-sufficiency of natural selection: A reply to Herbert Spencer. *Contemporary Review*, *64*, 309–338.

Yang, J., Lee, S. H., Goddard, M. E., & Visscher, P. M. (2011). GCTA: A tool for genome-wide complex trait analysis. *American Journal of Human Genetics*, *88*, 76–82.

Yang, J., Manolio, T. A., Pasquale, L. R., Boerwinkle, E., Caporaso, N., Cunningham, J. M., et al. (2011). Genome partitioning of genetic variation for complex traits using common SNPs. *Nature Genetics*, *43*, 519–525.

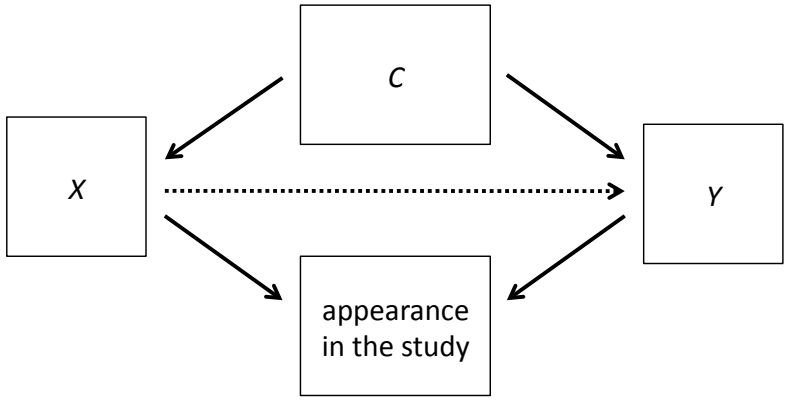
Author Note

Please send correspondence to jameslee@wjh.harvard.edu.

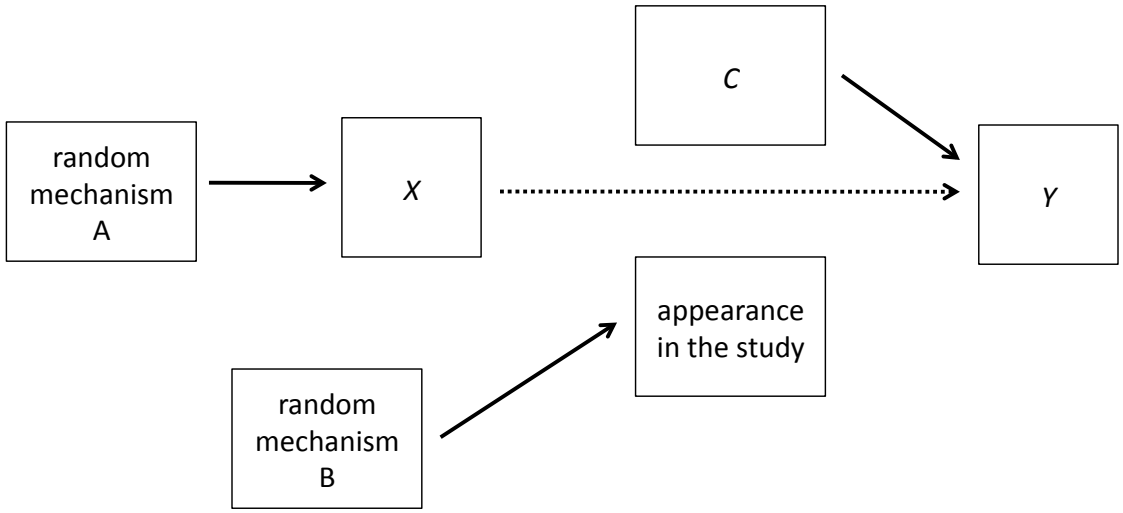
Figure Captions

Figure 1. DAGs representing the distinction between confounding and selection bias. (a) The case of no *a priori* assumptions regarding the relation between X and Y. (b) Randomization of X deletes the $C \rightarrow X$ edge. Randomization of study appearance deletes both $X \rightarrow$ *appearance in the study* and $Y \rightarrow$ *appearance in the study*. If X and Y are still associated (d-connected), it must be because of $X \rightarrow Y$.

(a)


C
X
Y
appearance
in the study

(b)


C
random
mechanism
A
X
Y
appearance
in the study
random
mechanism
B