European Journal of Personality, *Eur. J. Pers.* **26**: 391–413 (2012)
Published online in Wiley Online Library (wileyonlinelibrary.com) **DOI**: 10.1002/per.1865

OPEN PEER COMMENTARY

Bias due to Controlling a Collider: A Potentially Important Issue for Personality Research

JENS B. ASENDORPF
Department of Psychology, Humboldt University Berlin, Germany
jens.asendorpf@online.de

Abstract: I focus on one bias in correlational studies that has been rarely recognised because of the current taboo on discussions of causality in these studies: bias due to controlling a collider. It cannot only induce artificial correlations between statistically independent predictors but also suppress or hide real correlations between predictors. If the collider is related to selective sampling, a particularly nasty bias results. Bias due to controlling a collider may be as important as bias due to a suppressor effect. Copyright © 2012 John Wiley & Sons, Ltd.

In his stimulating paper that is unfortunately sometimes hard to read, Lee (this issue) touches a taboo topic in current personality publications: causal relations among variables that describe between-person differences. For many years, authors were educated by reviewers and editors to avoid causal language because of the many pitfalls in causal interpretations of correlations. These pitfalls granted dismissing causality altogether are like throwing out the baby with the bathwater. As humans, we cannot avoid thinking in terms of causality, and therefore, tabooing this topic in publications does not prevent readers and mass media from their own causal interpretations guided by implicit rules such as 'A correlates with B' means 'A causes B' but 'B correlates with A' means 'B causes A'.

Although causality is a difficult concept in correlational studies, scientists should and actually can do better than this if they can be pressed to explicate the causal model, or alternative causal models, underlying their research questions. The directed acyclic graph (DAG) method described by Lee (this issue) is a valuable method of achieving such an explication (see Foster, 2010, for an excellent discussion of causality based on DAGs for developmental psychologists). My comment here focuses on a key concept in the DAG approach: the collider.

Bias due to explicit control of a collider: Example from research on adaptation

A collider is an outcome of two joint predictors that may be correlated or not. If one statistically controls for a collider, the resulting correlation between the predictors will be necessarily biassed. Although this bias is most often discussed only for the case where two predictors are uncorrelated such that the bias consists of a spurious correlation, the bias is in fact general: any correlation will be biassed by the adjustment. As Lee (this issue) has correctly observed, the bias is obvious but rarely noticed by researchers.

For an example, let us consider data on risks and resources for adaptation of immigrant youth in Greece to the Greek culture (Motti-Stefanidi, Asendorpf, & Masten, 2012). Self-efficacy is an important resource, so the association of immigrant status with self-efficacy provides important information. Do these immigrants have lower self-efficacy expectations than their Greek peers? The answer is yes (the zero-order correlation between dummy-coded immigrant status in a sample of 969 adolescent immigrant students along with their Greek classmates was $-.15$, $p < .001$).

In studies of immigrant adaptation, skills in the host language are often routinely controlled because they may already explain most or all effects of other predictors of adaptation (although in many cases, suppressor effects may occur because the effect of language skills on adaptation is relatively strong). In the aforementioned case, if one controls the correlation between immigrant status and self-efficacy for the ability to speak Greek, the resulting partial correlation is $-.03$ and not significant any more. The control of Greek speaking skills induces a bias due to a collider because these skills are very likely causally influenced by both immigrant status and self-efficacy. Indeed, the respective correlations were $-.37$, $p < .001$ and .23, $p < .001$. Thus, controlling for host language skills is highly problematic in studies of the adaptation of immigrants where a resource and/or an adaptation outcome influence these skills because in such cases one controls a collider.

If one starts with explicit causal models *before* decisions are made on the statistical control of variables in the model, one will rarely commit this kind of erroneous over control. But if no causal analysis is made and the models involve many variables, or variables where it is not clear whether they should be considered a predictor or an outcome, researchers can easily be *lost in covariation*, relying on traditional routines designed for the control of certain predictor variables although they might be outcomes in the present context.

Bias due to implicitly controlling a collider through sampling: Example from research on achievement

If a collider is related to sampling such that the sample of participants is restricted in variation on the collider, this is equivalent to statistically controlling part of the variation of

the collider and therefore also introduces a bias. This is a particularly nasty case because the researcher did not explicitly control for the collider—the collider was implicitly controlled through selective sampling.

A surprising finding from research on achievement may illustrate this bias (I am grateful to Marco Perugini who alerted me to this case). Studies that relate IQ and conscientiousness to achievement regularly find the expected positive correlations of IQ and conscientiousness with achievement but at the same time nonsignificant or even negative correlations between IQ and conscientiousness (see, e.g. the meta-analysis by Ackerman & Heggestad, 1997), and authors who tried to explain this unexpected result had difficulty finding *post hoc* explanations. A causal analysis suggests a bias introduced by controlling a collider related to sampling. Most of these studies used university students or samples biassed toward high achievement, and this bias in sampling alone induced a negatively biassed correlation because achievement is a collider of IQ and conscientiousness.

Note that this bias is different from effects of restricted variance that can inflate or suppress a correlation but cannot induce a spurious correlation or change signs of a correlation. Bias due to implicit control of a collider can do this and may be even more common than bias due to explicit control of a collider. Many personality researchers are aware of biases due to restricted sampling, and the possibility to correct for them, but there is no tradition to consider biases due to the implicit control of a collider.

Importance for personality research

Right now it is hard to judge the importance of biases in personality research that are due to explicit or implicit control of colliders because both biases are largely unexplored. For the time being, a working hypothesis is that they may be as important as the better known suppressor effects between multiple predictors of the same outcome.

What Kind of Causal Modelling Approach Does Personality Research Need?

DENNY BORSBOOM[1], SOPHIE VAN DER SLUIS[1,2], ARJEN NOORDHOF[1], MARIEKE WICHERS[3], NICOLE GESCHWIND[3], STEVEN H. AGGEN[4,5], KENNETH S. KENDLER[4,5] AND ANGÉLIQUE O. J. CRAMER[1]

[1] *Department of Psychology, University of Amsterdam*

[2] *Complex Trait Genetics, Department of Functional Genomics and Department of Clinical Genetics, Centre for Neurogenomics and Cognitive Research (CNCR), FALW-VUA, Neuroscience Campus Amsterdam, VU University Medical Centre (VUmc)*

[3] *European Graduate School for Neuroscience, SEARCH, Department of Psychiatry and Psychology, Maastricht University Medical Centre*

[4] *Virginia Institute for Psychiatric and Behavioural Genetics*

[5] *Department of Psychiatry, Virginia Commonwealth University*

dennyborsboom@gmail.com

Abstract: Lee (2012) proposes that personality research should utilise recent theories of causality. Although we agree that such theories are important, we also note that their empirical application has not been very successful to date. The reason may be that psychological systems are frequently characterised by feedback, nonlinearity and individual differences in causal structure. Such features do not preclude the application of causal modelling but do limit the usefulness of the approach for the analysis of typical personality data. To adequately investigate personality, intensive time series of repeated measurements are needed. Copyright © 2012 John Wiley & Sons, Ltd.

We agree with Lee that recent theories of causality (Pearl, 2000) are important additions to the methodological literature and deserve more widespread study in psychology. In addition, as will be clear from our own paper (this issue), Lee's conceptualisation of personality in terms of (causal) networks is closely related to our own. However, although the theoretical value of Pearl's work stands beyond doubt, it is not as clear that the methodology is invariably fruitful in psychological applications.

In our experience with the kind of analysis that Lee attempts in his empirical data example, such analyses often fail to return readily interpretable results. Lee's analysis suffers the same fate, and its illustrative value is thereby limited to showing that there is a problem in either the causal assumptions or the data, or both. This prompts the question whether the data currently at our disposal, which typically concern individual differences at a single time point, are adequately suited for the application of causal modelling in the context of personality. In our view, there are two aspects of psychological systems in general, and personality research in particular, that hamper causal modelling of the data we typically have in personality research. First, personality processes are likely to involve feedback. Feedback is optimally analysed in time series, which are not often available. Second, when applied to one-shot test data (i.e. when a large number of people answer questionnaire items at one specific

DOI: 10.1002/per

moment in time), causal modelling requires that the system analysed is invariant over the units of observation; that is in psychology, it requires that people are homogeneous in their organisational and causal structure. This is, in our view, unlikely (see also Molenaar, 2004).

Regarding the first point, although there have been some stabs at modelling cycles in graphs, the main power of the graphical approach to causality lies in the analysis of directed acyclic graphs. Note that directed acyclic graphs may be used to model feedback (e.g. Dahlhaus & Eichler, 2003), but that is not possible if the data are not sampled from the time domain. However, feedback is likely to play a very important structural role in developing and sustaining personality. As we argued in our own paper (this issue), feedback is likely to operate through the selection of situations where the individual can express behaviour that is in some sense rewarding. However, it also pops up in the basic regulation of the body through homeostatic couplings between properties (e.g. sleep and fatigue, mood and activity, etc.). It is likely that such feedback processes, which are instrumental in controlling the most fundamental aspects of human functioning, are also important in shaping and maintaining personality features.

In many practical cases, the causal modelling approach (which is in principle nonparametric in nature) is not only limited to the analysis of directed acyclic graphs but also supplemented with linearity assumptions (a primary and well-known example of this involves typical applications of structural equation modelling). However, in the case of homeostasis, we almost always see variables that are bounded from below and from above (e.g. one cannot sleep less than zero or more than 24 h a day, eat less than nothing, etc.), which means that relations between such variables cannot be linear. Thus, feedback must be coupled with nonlinearity. It is not clear that if one samples the result of such processes at a single time point across individuals (the typical data-gathering setup in personality research), anything reasonable and coherent can be expected from applying standard causal modelling using directed graphs.

Second, the approach that Lee utilises is predicated on the assumption that the people studied can be described by the 'same' causal organisation. In personality, it is not clear that this assumption is viable. For instance, some people react to stress by eating less; some by eating more. Some people become agitated when they lose sleep; some become slow. Some people react to fear by fighting, some by fleeing and others by freezing. It is not even established that such patterns remain stable over development. Although such individual differences do not rule out the application of causal models if these are fitted to individual time series (e.g. Dahlhaus & Eichler, 2003; Eichler, 2007), it does call into question the patterns of conditional independencies that are required for the sort of individual differences data that Lee analyses.

Both feedback and individual differences can be addressed within the causal framework of Pearl (2000) if one is willing to invest in gathering sufficiently long time series that can be analysed at the level of the individual. Such data are becoming more frequent, and it is to be hoped that researchers are willing and able to make a large-scale transition in its research practice towards gathering such data. However, we live in a time when many researchers still interpret personality traits as forces that operate at the level of the individual person, which leads to the inaccurate impression that personality psychology is already studying the mechanisms of individual behaviour. Lee does not commit to this interpretation. However, although we agree that factor models need not necessarily be causally interpreted in all applications, it is unclear to us what his alternative conceptualisation of factors comes down to. We note that the application of the theory of McDonald (2003), which Lee invokes, relies on the interpretation of factor scores as scores on infinite unidimensional item domains (tail measures; see Ellis & Junker, 1997; Markus & Borsboom, 2011). It is, in our view, unlikely that this interpretation is satisfactory in personality theory, which does not even have finite unidimensional item domains.

In conclusion, we agree with Lee that recent theories of causality deserve more widespread dissemination. However, it is still unclear whether the empirical methods associated with these theories will bear fruit in the area of personality research. In our view, it is more likely that a redirection of research towards the study of the dynamic structure of individuals will better move the field of personality research forward.

Scale Issues in Causality

DAVID M. CONDON, ASHLEY BROWN-RIDDELL, JOSHUA WILT AND WILLIAM REVELLE

Northwestern University

revelle@northwestern.edu

Abstract: Elaboration of the manner by which graphical frameworks of causality can benefit personality research is a much-needed contribution. We argue that attempts to apply these frameworks in personality will benefit from consideration of two concepts related to scale. The first is that the appropriate scale on which to evaluate causality depends

DOI: 10.1002/per

upon the level of analyses on which the research is conducted. Second, the distal scale between typical expressions of personality and their possible causes limits discussion of causality to probabilistic rather than mechanistic factors.

By virtue of even attempting to integrate Judea Pearl's innovative work on causality (Pearl, 2009) into the personality psychology literature, James Lee is to be commended for initiating a conversation that is—in truth—uncomfortable for a field of 'correlators' (Cronbach, 1957). But it is Lee's concise elaboration of so many nuances of the graphical framework that makes the target article an invaluable contribution to the field, particularly for those unfamiliar with Pearl's work. To the extent that personality psychologists increasingly focus on the development and evaluation of predictive causal models, we consider it likely that the influence of this work by Lee will grow over time. Although many aspects of Lee's review merit further exploration, our commentary primarily focuses on two aspects of the relationship between the causal framework that Lee describes and considerations of scale.

The first point of note pertains to Lee's comments on the role of psychometric factors in graphical conceptualisations of psychological causality. We agree with Lee that one of the goals of personality research should be to recursively 'expand a directed edge in one graph into an entirely new subgraph' (p. 387) to understand the mechanistic relationship between two nodes, an endeavour that would ideally result in richly detailed causal diagrams similar to those found in biology texts. Achieving this level of mechanistic detail may well resolve debates about the causal status for many psychometric factors, but it would not mitigate the functional utility that factors provide.

By analogy, a graph is something like an online geographic map. Psychometric factors, like many cartographic features, are not physically observable; researchers would no more benefit from 'discarding the convenient fictions of folk psychology' (p. 38) than travellers would if the town, state and nation labels were stripped from maps. The familiar experience of 'zooming' an online map to the appropriate viewing level illustrates the contextual utility of complexity, which is itself a function of scale. Practically, infinite detail is possible in mapping of both human behaviour and geography, but this would rarely be functional. Conceptual relationships are more easily grasped and manipulated when they are manageable in number and roughly similar in scale.

For example, it is common for trait psychologists to lament the broad imprecision of factors like the Big Five that are borne out of data reduction, yet most would agree that this is the appropriate level of analysis for evaluating topics such as the differential relationships on career outcomes of extraversion and cognitive ability (Roberts, Kuncel, Shiner, Caspi, & Goldberg, 2007). More narrow organisational frameworks such as the Big Five Aspect Scales (DeYoung, Quilty, & Peterson, 2007) or the 30 NEO facets (Costa & McCrae, 1992) would be appropriate for exploring the differential relationships of various facets of extraversion. An example of zooming in still further might include recent theoretical work relating an individual's desire to engage in prototypically extraverted behaviours to physiological measures of dopamine (Smilie, Cooper, Wilt, & Revelle, in press).

'Fictional factors' play an integral role at each of these levels of analysis in the form of latent behavioural trait constructs. Admittedly imprecise at all levels, factors enable researchers to delineate continuous streams of observable behaviours into chunks that can be reasonably well measured and organised. The speculative proposition to map each of the hypothetical relationships between low-level biology and long-term, multiply determined outcomes such as extraverted behaviours that affect career outcomes would seem to neglect this tension between precision and efficacy. To be clear, we do not argue against the merits of model specificity but rather for the merits of appropriate model scaling. Whereas it is invaluable to be reminded that the factor analytic approach can only inform a subset of the questions personality psychologists hope to explore, it is also true that the parsimonious chunking of data that factoring allows will likely justify its continued application. Lee stops short of offering psychological researchers a metric for identifying and pursuing the appropriate levels of analysis.

Our second point is to emphasise that the scale on which personality unfolds requires acknowledgement of the distinction between mechanistic and probabilistic causality. Intuitively speaking, the complexity of pathway navigation is a function of proximity. To be more precise, the number of intermediate nodes is the variable that determines the number of alternative pathways, but it is effectively true that outcomes that are temporally or physically proximal to their causes are predicted with higher probability than those that are distal. In other words, the pursuit of mechanistic causality is a reasonable aspiration on a small scale because outcomes and their purported causes are relatively proximal.

Most outcomes under study by personality psychologists, however, are multiply determined over long periods of time. This suggests that the majority of causal factors are probabilistic rather than mechanistic. Conditioning discussions of causality on this distinction is vital as the very basis for researching personality and cognitive ability would be altered if these constructs were found to be mechanistically determined by genetics. The distinction is also sobering in that it forces us to acknowledge that personality is determined by the cumulative influence of thousands of genes, a nearly infinite variety of environmental variables, and multiple types of gene by environment interplay (Johnson, 2007; some of which may occur more commonly than the rare incident of mutation claimed by Lee as justification for general temporal restriction [p. 40]). These genetic and environmental inputs do not

DOI: 10.1002/per

directly cause behaviour but are rather mediated through their effect on proteins and subsequent neural systems that lead to differential environmental sensitivities resulting in different cognitive, affective and motivational values.

Lee acknowledges this complexity and clearly explains why the proximity of cause and outcome is less relevant for gene-trait association studies. However, this does not imply mechanistic causality for the expression of the trait in any given context. Unlike the examples made of height, hair morphology and Parkinson's, personality constructs describe typical manners of behaving across a wide variety of contexts. We take this to mean that future knowledge regarding gene-trait associations would only allow for probabilistic estimates, for example, of a given individual's typical desire to attend lively parties as an opportunity to express the extraverted tendencies that result from the dopaminergic effects on their wanting system (Smilie et al.,). Although this type of knowledge will someday constitute an impressive contribution to the field, it will not reflect 'causality' as traditionally defined.

The Wright Stuff: Genes in the Interrogation of Correlation and Causation

GEORGE DAVEY SMITH

MRC Centre for Causal Analyses in Translational Epidemiology, University of Bristol

KZ.Davey-Smith@bristol.ac.uk

Abstract: The contemporary use of what are now called causal diagrams can be traced back to Sewall Wright's introduction of path coefficients in the early 1920s. Wright was explicit that causal evidence was required to formulate such diagrams and that these schema could not alone provide evidence for or against causality. In population sciences, germline genetic variation can provide required anchors for the separation of causal from (mere) correlational associations. Advances in biological and other material sciences offer more for improved causal understanding than new ways of conceptualising and representing associations. Copyright © 2012 John Wiley & Sons, Ltd.

For most branches of science, the distinction between (mere) correlation and causation is a central issue. My discipline, epidemiology, is one prone to over interpretation—by the media, by researchers or both—of associations observed in data sets that are most plausibly explained by chance, bias or confounding (Davey Smith & Ebrahim, 2002). James Lee muses on these issues in the context of behavioural traits within the psychological literature and promotes the graphical approach (in particular, directed acyclic graphs) now beloved of many working within the epidemiological tradition. His clear presentation merits a close reading and raises issues of general relevance. I will focus on the opportunities offered by his statement that '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.' Indeed, leveraging the power of germline genetic variation transforms our ability to elucidate the causal chains within the networks of associations within the biological realm (Zhu et al., 2007), and whereas graphical presentations may help, it is the biological realities, rather than new ways to draw these on paper, that contain the most promise. These are only now beginning to yield findings but will transform how we approach causality in the population sciences.

Lee invokes the evolutionary biologist and population geneticist Sewall Wright, the progenitor of path analysis (and, through that, structural equation modelling, favored more in the psychological than epidemiological literature) in the prehistory of the now triumphant directed acyclic graph. I must admit to being pleased that structural equation models largely failed to penetrate epidemiology; their (sometimes) manner of presentation as a form of alchemy that can isolate causal pathways in an intercorrelated morass of data being scarcely credible. In the epidemiological setting, underlying social and biological processes, combined with reverse causation (outcome influencing apparent exposure, rather than *vice versa*), leads to association being the norm rather than the exception (Davey Smith et al., 2008). Levels of measurement error that exist in most domains simply cannot be disciplined, and the confident production of coefficients that apparently have meaning seems chimeral. Thus, coming across Wright, authoring a paper in 1921 with the exact same title as Lee, setting out his stall for his form of path analysis was enlightening:

> The ideal method of science is the study of the direct influence of one condition on another in experiments in which all other possible causes of variation are eliminated. Unfortunately, causes of variation often seem to be beyond control. In the biological sciences, especially, one often has to deal with a group of characteristics or conditions which are correlated because of a complex of interacting, uncontrollable, and often obscure causes. The degree of correlation between two variables can be calculated by well-known methods, but when it is found it gives merely the resultant of all connecting paths of influence.
>
> The present paper is an attempt to present a method of measuring the direct influence along each separate path

DOI: 10.1002/per

> in such a system and thus of finding the degree of which variation of a given effect is determined by each particular cause. The method depends on the combination of knowledge of the degrees of correlation among the variables in a system with such knowledge as may be possessed of the causal relations. In cases in which the causal relations are uncertain the method can be used to find the logical consequences of any particular hypothesis in regard to them. (Wright, 1921, p. 557)

Wright's famous path analyses (Figure; Wright, 1920) required prior causal knowledge to make sense. With this, they introduced important new understanding, not the least of which was the identification of what Wright termed 'intangible variance'—induced by what we may call stochastic or chance events—that lead to group level, rather than individual trajectory, understanding. This is the best that can ever be hoped for in the population sciences (Davey Smith, 2011b).

To an extent, known biological relationships in quantitative genetic analyses in the behavioural genetics field provide a form of reliable prior information on the presence and direction of causation. However, in the molecular genetics era, the most powerful source of prior causal knowledge that can, and is, now being leveraged comes from germline genetic variants that have established associations with particular traits. R.A. Fisher explicitly referred to the essentially randomised nature of genetic perturbations (Fisher, 1952), as Lee mentions and as others directly associated with Fisher have written about (Bodmer, 2003; Box, 2010), although the possibility that Mendelian randomisation came before experimental randomisation in Fisher's intellectual biography has been little recognised (Davey Smith, 2006). That genetic variation inducing a group-level difference in a potentially modifiable phenotype can provide evidence of the downstream causal effect of this phenotype, free of the influence of confounding or reverse causation, is now widely recognised and implemented in epidemiological studies (Timpson, Wade, & Davey Smith, 2012). To give just one example of relevance to the study of behavioural traits—the topic of Lee's paper—such 'Mendelian randomization' (as the method is generally termed; Davey Smith & Ebrahim, 2003) has been applied to the effects of smoking. As proof of principle, such studies have demonstrated that a genetic variant robustly associated with smoking behaviour relates to lung cancer risk to the degree expected by the association of the variant with appropriately ascertained smoking behaviour (Munafo et al., 2012; Wang, Broderick, Matakidou, Eisen, & Houlston, 2011), and associations with several other smoking-related diseases have been made. Such studies have also shown that smoking lowers body mass index (Freathy et al., 2011); despite naive observational associations sometimes being in the opposite direction, given confounding by socioeconomic position and various other socially patterned exposures.

The various assumptions of such Mendelian randomisation studies have been reviewed (Davey Smith, 2010; Lee, this issue; Sheehan, Didelez, Burton, & Tobin, 2008) and are reflected in Lee's discussion of the distinction between Fisher's notion of the as-observed 'average excess' associated with a genetic difference and the 'average effect' that would be seen with a gene substitution. That confounding can exist in genetic association studies is of course widely recognised, with ancestral population differences in both gene frequency and disease risk ('population stratification') being the most likely culprit. There are well-established methods of accounting for this using genome-wide data as indices of such population stratification, and with established genetic variant-phenotype links, it is remarkable how homogeneous the associations seen within different populations generally are, despite allele frequency often varying between populations (Hindorff et al., 2012). Empirical data also demonstrate that confounding of genetic variants with social, behavioural and physiological factors that plague conventional observational studies are conspicuous by their absence (Davey Smith et al., 2008).

Lee considers at length the possibility that selection bias related to participation in studies could bias findings. Thus, if a genetically influenced trait was related to willingness to participate in a study, and this was differential for cases and controls, a spurious association could be generated. This is in principle true, but common control groups have been used for various diseases (e.g. the Wellcome Trust Case Control Consortium, 2007), and unless the participation effect was condition specific, such bias would generate similar associations for all the diseases, which were not seen. Even if such a participation effect was disease specific, it would only influence case-control studies, not prospective studies, and generally, genetic associations have been similar across study designs (Hindorff et al., 2012). More complex hypotheses could be advanced involving interactions of genetic variants influencing participation and condition-specific disease risk, but plausibility decreases with increasing elaboration of the hypothesis in this regard. Again, the fact that similar effects for established variants tend to be seen in designs with widely differing participation rates, from high response rate general population cohorts to what are essentially volunteer studies, is reassuring in this regard.

Graphical approaches to causal inference are certainly of value in forcing investigators to be explicit about their assumptions and can help in the identification of unrecognised potential biases. There are also often unrecognised drawbacks to formulaic or mechanical imposition of such approaches (Dawid, 2008). In epidemiological circles, it is now not uncommon to receive peer review comments that focus on "the possible adjustment for a collider in model 3 of Supplementary Table 4. The reviewer clearly considering this more important than having an informed overview of the totality of evidence presented. 'Inference to the best explanation' (Lipton, 2004), which is surely what any attempt at causal reasoning is aiming at, can go out of the window as the d-connected nodes, rather than how the world actually is, become the focus of attention.

DOI: 10.1002/per

Wright opined that 'great refinement in statistical treatment is often a waste of effort' (Wright, 1917). William Provine (1986), in his unsurpassable intellectual biography of Wright, discusses the development of path analysis and how, working with methods that tried to hold other factors constant through statistical manipulation, 'Wright was still dissatisfied. He saw clearly that by itself the partial correlation coefficient, like the correlation coefficient, was a mathematical quantity not tied or leading by itself to any causal interpretation of the relations under examination. Wright wanted to minimise correlational statistics and maximise the quantitative causal interpretation of the variables' (p. 127). This can only be carried out when causal anchors—that come from how the material world is, not how we draw diagrams on paper—are introduced into the mix. Germline genetic variants provide precisely such anchors and open up vast new vistas of possible causal understanding generated from observational data (Davey Smith, 2011a).

ACKNOWLEDGEMENTS

Thanks to Dave Evans for discussion of selection bias in genome-wide association study.

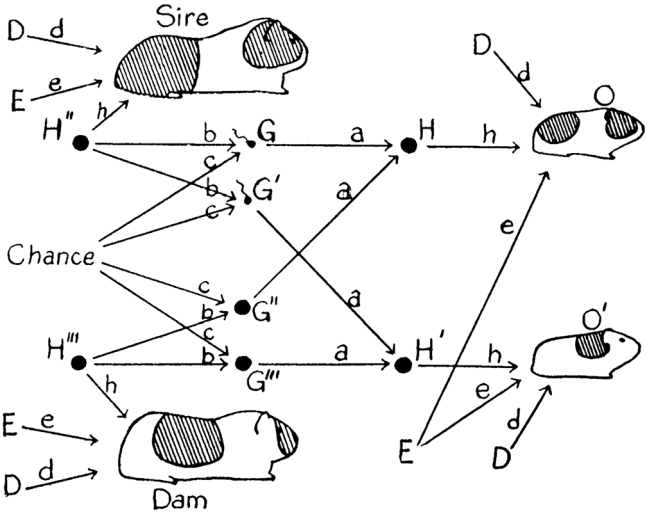

Diagram illustrating the casual relations between litter mates (O, O') and between each of them and their parents. H, H', H" and H"' represent the genetic constitutions of the four individuals; G, G', G" and G"' represent that of four germ cells. E represents such environmental factors as are common to litter mates. D represents other factors, largely ontogenetic irregularity. The small letters stand for the various path coefficients.

Causing a Shift in Causal Thinking

JOSHUA J. JACKSON[1] AND SETH M. SPAIN[2]

[1]*Washington University in St. Louis*
[2]*Binghamton University and State University of New York*

j.jackson@wustl.edu

Abstract: We concur that the difficulties of causal analysis are especially problematic for personality psychologists. We suggest that this stems from both historical and methodological reasons. Additionally, we note that the Pearl model is not completely adequate for some questions pertinent to personality psychologists and mention the existence of underutilised methods that provide stronger causal inference. It is important to remember that many hypotheses are causal in nature and that designs other than randomised experiments exist to estimate causal effects. However, no single design can guarantee the identification of causality. Copyright © 2012 John Wiley & Sons, Ltd.

Discussions of causality within personality psychology are usually relegated to a research methods class or the occasional methods chapter and rarely (ever?) make their way into journal introductions or discussion sections. Of course, this is despite many of our hypotheses being *causal hypotheses* (Pearl, 2000). Most likely, this is due to personality psychologists knowing that each research design we employ is imperfect and, thus, incapable of producing strong causal conclusions. Instead of openly discussing causal hypotheses, it is easiest to err on the side of caution. Therefore, we omit direct reference to causal relationships and instead speak in safe terms of 'associations' or 'prediction'. One of the highlights of the current article is to appreciate that we are often interested in causal questions. If personality psychologists desire to provide input to policy, wrestling with issues of causality is a necessary task (Almlund et al., 2011). Moreover, in the interest of conducting better science, it is important to be open about causal hypotheses as our theories drive the way in which we conduct and analyse our research.

How do we think about causality?

Currently, psychologists (and other related fields) tend to have a relatively narrow and unified view of causality (White, 1990). This causal framework was built out of the philosophical discussion of causality first championed by Hume and Mill and then developed into formal statistical models. This model, the Rubin model of causality (proposed by Holland but based on ideas introduced by Neyer and Rubin, thus also called the Neyer–Rubin or Holland–Rubin

DOI: 10.1002/per

model; Holland, 1986; Rubin, 2005) focuses on providing a framework to make the strongest causal inference possible, whose 'gold standard' is the randomised experimental design (Campbell & Stanley, 1963). Unfortunately, the observational and correlational designs employed by personality psychologists do not usually fit this model given that individual differences are not readily manipulable, and certain questions necessitate observational or correlational designs because of ethical, temporal or financial reasons. Given this purview, causality is not discussed when these less than optimal designs are employed. Rather, a cautious and agnostic approach is taken in this instance because the design is not the gold standard randomised experimental design. We believe that this prevailing model of causality has become so prevalent that what is causal is solely defined by the method endorsed by this model. In other words, it is assumed one cannot find causal associations unless a manipulation occurs (Holland, 1986).

This purview is unfortunate because not only it limits our ability to speak in causal terms but also because there exist other approaches to think about causality and test causal relationships (Goldthorpe, 2001). In psychology, structural equation modelling is often brought up as an obvious candidate. Of course, the structural models are simply covariance matrices that in and of themselves do not necessarily capture many of the agreed upon requirements for causality (Bollen & Pearl, in press). The Pearl model nicely builds upon this framework and provides a way to think about causality that does not necessarily rely on manipulation of variables. The positives of this model are many, but a few difficulties exist, especially for personality psychologists.

One major challenge for the Pearl causal approach is that it cannot, generally, handle issues of reciprocal causation or, more specifically, nonrecursive models. Within Pearl's approach, causal models must take the form of *directed acyclic graphs*. So, $X \rightarrow Y$ or $X \leftarrow Y$, but $X \leftarrow\rightarrow Y$ is not generally allowed (where $\leftarrow\rightarrow$ represents mutual causation, not correlation). Cycles, such as reciprocal causation, will typically result in a breakdown in the mathematical logic that leads to causal knowledge. Within personality psychology, many different theories posit that two variables may affect each other across time and thus would not be applicable to the type of analysis presented in the current paper. Other causal modelling traditions directly deal with these difficulties, such as some econometric models (Heckman, 2005).

How can we test causal relationships?

As mentioned before, we believe that a number of designs other than randomised experiments can provide causal inferences. These designs do so by safeguarding against many biases that arise in observational studies. One common technique is to use heredity to our advantage (e.g. Davey Smith & Ebrahim, 2003; Rutter, 2007). Two relatively new—for psychologists—methods of causal analysis are propensity score matching and instrumental variables (IVs).

Because observational designs are not randomised, selection biases exist. As a result, it is necessary to control for any confounding pre-existing differences that could bias the estimation of causal effect. Typically, this is carried out through the inclusion of covariates in a standard linear model. In contrast, propensity score matching offers a much stronger way to control for selection bias in observational data by simultaneously controlling for many covariates (Thoemmes & Kim, 2011). In this approach, each participant receives an estimated propensity score, which is the conditional probability that a given participant would be exposed to the treatment condition, given certain values on observed covariates. By matching participants that have or have not been exposed to the treatment on this estimated propensity score, pairs of participants are created that are balanced on all observed covariates—a situation that would be expected under a randomised experiment. This matching process creates two distributions that are balanced with regard to observed background variables that may bias our findings. Accordingly, these two distributions only differ in terms of the treatment that they received. This approach is slowly being integrated into psychology (e.g. Jackson et al., 2012; Kendler & Gardner, 2010) but has yet to fully utilised.

Another major threat to causal inference is endogeneity (Antonakis, Bendahan, Jacquart, & Lalive, 2010). Endogeneity occurs when a predictor or set of predictors is correlated with the error term in a model (Kennedy, 1985). Endogeneity is especially problematic for personality researchers because it can occur when important variables are omitted from the model or if there is reciprocal causation between outcome and predictor(s). The result of endogeneity is a biassed estimate of model coefficients.

Instrumental variables are one method to safeguard against endogeneity. An IV is a variable correlated with the regressor but independent of the error term. Two-stage least squares is used to estimate the effect of a 'purified' regressor, cleansed of its error-correlated variance. Say we have a model of the type, $Y = b_0 + b_1X + \varepsilon$, where $\text{cov}(X,\varepsilon) \neq 0$. In such a case, the estimate of b_1 will not be equal to the 'true' population value and is thus a biassed estimate. IV methods are used to 'clean' the variance of X association with ε. In essence, IV methods first regress X on an instrument, Z, to obtain a purified X^*, and then Y is regressed on X^* to obtain an unbiased estimate (Gelman & Hill, 2007).

For example, consider the hypothesis that extra schooling increases IQ. A policy change that would lead to a greater number of years of education, such as some states providing a large number of scholarships, could serve as an instrument, provided that scholarships were not based on IQ. This policy change would not directly affect the outcome, IQ, and thus could act as an instrument to better test the causal effect of whether schooling leads to increases in IQ.

Summary

It must be remembered that no single design can definitively identify a causal relationship. For all of the positives of randomised experiments, there exist shortcomings that do not guarantee causal inference. In the end, the best way to understand causality is by growing a nomological net—testing and retesting, ruling out alternative hypotheses and replicating findings. Personality psychologists, however, can do this as well as any other discipline and thus should not be coy about their causal hypotheses.

DOI: 10.1002/per

Petard

WENDY JOHNSON

Department of Psychology, University of Edinburgh
Centre for Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh, Edinburgh, UK

wendy.johnson@ed.ac.uk

Abstract: Lee has made an important contribution in introducing personality psychologists to Judea Pearl's rigorous conceptual approach to thinking about causality. The approach makes clear the dangers in applying commonly used statistical controls without considering all alternative causal mechanisms. Ironically, Lee chose gene-trait association as a likely ultimately causal model. He was likely hoist with his own petard when he omitted the path from environmental circumstances to gene expression. No abstract conceptual approach can be a complete substitute for clear and objective thought about alternative explanations. Copyright © 2012 John Wiley & Sons, Ltd.

In warfare, a siege takes place when an army attacks a fortress that cannot easily be taken and refuses to withdraw. Personality psychology has long been engaged in a siege against the fortress of complexity surrounding the emergence and application of personality traits in the lives of individuals. The siege's progress ebbs and flows: occasionally, we stumble on a crack in the fortress walls through which we can hurl a bomb, but often, it seems that the better our weapons, the higher those walls grow. Lee, however, has introduced personality psychologists to an important new weapon in the elegant and revealing conceptual approach to causality of Pearl (2009).

Pearl's weapon reminds me of the petard. In mediaeval and Renaissance siege warfare, a petard was a small bomb. It was used to breach fortifications by blowing up specific gates or walls. Petards were often placed inside tunnels under walls or directly upon gates. Like Pearl's approach to understanding causality, when it worked, the petard was a sophisticated, targeted tool with considerable power to accomplish a much larger goal. Also like Pearl's approach, however, successful use of the petard meant keeping track of the interconnections among many different contributing factors. For example, it was common for the attacker to dig a shallow trench close to the fortress gate. The engineer would then erect a small hoisting engine of wood, ropes and pulleys from this trench. The engine was designed to lift the petard, once lit by the engineer, out of the trench and to hurl it at the gate, where it would detonate, destroying the gate. Sometimes, however, in activating the system, the engineer would become entangled in the ropes and hoisted out along with the petard and blown up himself, 'hoist by his own petard'. In *Hamlet*, Shakespeare used this occurrence as a metaphor for the position in which Hamlet found himself: bearer of sealed letters from his Uncle Claudius to the King of England, instructing the King to execute him. Fans of Shakespeare have made the expression a commonly used figure of speech in English, meaning much the same as 'shooting oneself in the foot.'

Lee has made Pearl's approach relevant to personality psychologists through several apt examples, pointing out clearly the potential for misattributions about causality, especially when collider variables are statistically controlled. As he noted, such control can and too often does take place unintentionally through sample selectivity and also intentionally but unwittingly through application of statistical control that disregards alternative pathways. These examples play to the strength of the approach, which is to bring to light mistakes in causal attribution resulting from overly simplistic causal models. The approach can certainly be used actively to prevent such mistakes through examination of all possible alternatives, but the thoroughness of this process, and thus its effectiveness, lies completely in the hands of the researcher. The approach itself cannot ensure the necessary objectivity and independence of mind. Perhaps unintentionally, Lee made this instructively clear in Section 4. He selected gene-trait association as one of the most likely direct examples of ultimate causality. R. A. Fisher's model of additive genetic variance and heritability was considered a substantial contribution in its day, but it is becoming clearer by the moment that that pesky environment, broadly construed as circumstances both external and internal to the organism (including other genes in the genotype), repeatedly gets in the way, not only by making it hard to avoid genetically selected samples as Lee discussed but also by altering the very nature of gene expression.

And it is gene expression, well beyond gene presence in the genotype, that likely matters most for genetic influences on the patterns of behaviours that personality psychologists explore. To date, evidence for this is generally indirect in humans, and specific examples come primarily from model organisms (see Johnson, Penke, & Spinath, 2011 last year's target article in this journal and its associated comments for more complete discussion of this). Nevertheless, disregarding the relevance of these examples to human personality development is to trip over the guide rope just as you light and hoist the petard.

ACKNOWLEDGEMENTS

I thank Tom Booth and Rene Mottus for comments on an earlier draft.

DOI: 10.1002/per

Causality: Populations, Individuals and Assumptions

ROGIER A. KIEVIT[1], LOURENS J. WALDORP[1], KEES-JAN KAN[2] AND JELTE M. WICHERTS[1]

[1]*Faculty of Social and Behavioural Sciences, University of Amsterdam*
[2]*Biological Psychology Department, Free University of Amsterdam*

r.a.kievit@uva.nl

Abstract: We side with Lee on the importance and potential use of graphical causal modelling in personality research but raise three issues crucial to its validity. First, causal relations obtained at the inter-individual level should not be confused with intra-individual causal relations. Second, it is difficult to explicate all assumptions about which variables are measured, their causal relations and the possibility of co-occurring events when applying graphical modelling to personality data. Third, multiple testing complicates assessing (in)dependencies. Conclusions and inferences should always be drawn with appropriate attention to the underlying assumptions. Copyright © 2012 John Wiley & Sons, Ltd.

We agree with Lee (this issue) that the graphical framework as set out by Pearl (2009); Lauritzen (1996); Spirtes, Glymour, and Scheines (2000) and others provides a rich account of causality much needed in the social sciences. However, we would like to point out three issues in causal modelling that we believe were not given sufficient attention by Lee. First, causal relations obtained at the population level (inter-individual differences) are often applied to, and confused with, intra-individual causal relations. Second, to infer a causal relation, strong assumptions are required about which variables are measured, their causal relations and the possibility of co-occurring events. Third, when testing (in)dependencies, one is faced with the multiple testing issue.

In Lee's discussion, within-subject and between-subject analyses are often confounded. This is especially relevant for personality research, as personality dimensions are generally studied as inter-individual dimensions. Consider Lee's Figure 2, which is a description of a sequence of events unfolding over time under different settings. It entails a 'within-pavement' causal explanation. Imagine if instead of repeatedly observing this pavement over time (e.g. seasons), we were to observe the same indicators for different pavements across the world. This dataset will include pavements from rich and poor countries (where they can or cannot afford good sprinklers) with pavements in areas with 4, 3 or 2 distinct seasons, with differing building materials, and a host of variables that will affect their inter-pavement wet-becoming behaviour. Such a model will tell a different causal story, as it is concerned with differences *between* pavements. For instance, building materials may be a relevant factor in explaining the causal chain from raining to wetness *between* pavements but not *within* pavements.

These are two distinct causal explanations, and both are relevant for personality research. Although the conventional Big Five model represents individual differences, it can be extended to include intra-individual data, by actually measuring differences in behaviour and attitudes within people over time (Molenaar & Campbell, 2009). Structural equation modelling of time-series personality data reveals that the intra-individual pattern, and therefore the nature of the causal explanation, may be different within each individual that makes up some population, and therefore distinct from the inter-individual pattern (e.g. Hamaker, Nesselroade, & Molenaar, 2007). This confirms Lee's suggestion that a well-fitting model at the group level does not necessarily capture all causally relevant dynamics.

The distinction between within-subject and between-subject explanations is crucial (e.g. Penke et al., 2011), especially for biological explanations. In behaviour genetics and cognitive neuroscience, inter-individual findings are often interpreted as being causally operative at the level of the individual. For instance, as Figure 1 in Lee illustrates, the links between genes, brains and cognition are often assumed to be causal and one way. Recent studies suggest that psychological activities such as intense memory training can affect brain structure within individuals over time (e.g. Draganski et al., 2004). This can occur regardless of cross-sectional correlations between brain and behaviour. The biological phenomena that explain differences between people may differ from phenomena that explain changes within people over time. Or, in the do-syntax:$(grey\ matter|do(differences\ in\ memory\ performance)) \neq (grey\ matter|see\ differences\ in\ memory\ performance)$. Similarly, the heritability of certain psychological traits is a variable that explains differences between people, not the process (or inevitability of the outcome) within individuals. This distinction is essential yet often overlooked. Thus, we suggest that extra attention be paid to distinctions between inter-individual and intra-individual causal explanations and causal explanations that cross explanatory levels.

Our second point concerns the assumptions of estimating causal effects. Our view is that causal analyses are very powerful, but that assumptions should be made explicit so that researchers can properly evaluate the strength of the conclusions. Consider the example in Lee's Figure 3. To infer a causal effect of IQ on SES, we are required to use the back-door criterion that assumes that (i) all direct causes of IQ (i.e. have an arrow into IQ) are controlled for, (ii) no effects of IQ are controlled for and (iii) the probability of the direct causes of SES and IQ is almost nowhere zero.

Condition (i) implies that the set of measured variables is sufficient to block all paths to IQ. In general, this is

DOI: 10.1002/per

difficult to maintain or test. It is often unclear which variables are involved and/or there may be a lack of consensus about the direction of arrows. In the example, could education be a relevant variable, and if so, will it have a direct arrow into SES or will it be an effect of SES? If education is considered a cause of SES, then it should be controlled for; if education is considered an effect of SES, then it should not be controlled for. Such prior knowledge is crucial. The application of the formula of the back-door (Pearl, 2000) requires that the distribution is positive almost everywhere (Lauritzen, 1996). If this is not the case, the adjustment cannot be applied. In the example, condition (iii) entails that, for instance, there should be families with low parent SES and high offspring SES, and vice versa. Is this likely? And if these combinations do not co-occur, should we censor distributions? So assumptions of applying the back-door adjustment are quite strong and often difficult to test. We are in favour of graphical modelling, but users should be aware of the meaning of these crucial assumptions before applying such techniques.

A final issue concerns the analysis of complex graphical models that are often encountered in personality research. Consider a simple example with five nodes in a graph. To test all (in)dependencies implies that for each combination of nodes, we have six tests, and so for all $5*(5-1)/2=20$ combinations, we have 120 tests. This raises the question of how to deal with multiple comparisons. Should we use Bonferroni, False Discovery Rate (FDR) or local FDR?

Overall, Lee's arguments for the better incorporation of graphical models and causal inference are timely and well taken. However, for the aforementioned reasons, conclusions and inferences should always be drawn with appropriate attention to the underlying assumptions.

Correlation and Causation—The Logic of Co-habitation

JUDEA PEARL

Computer Science Department, University of California, Los Angeles

judea@cs.ucla.edu

Abstract: Recent advances in graphical models and the logic of causation have given rise to new ways in which scientists analyse cause–effect relationships. Today, we understand precisely the conditions under which causal relationships can be inferred from data, the assumptions and measurements needed for predicting the effect of interventions (e.g. treatments on recovery) and how retrospective counterfactuals (e.g. 'I should have done it differently') can be reasoned about algorithmically or derived from data. The paper provides a brief account of these developments. Copyright © 2012 John Wiley & Sons, Ltd.

James Lee's paper, Correlation and Causation, would probably raise a few eyebrows among readers of the *European Journal of Personality*. 'Again?' Some will ask, 'Haven't we heard enough about this subject? and isn't it an established fact that even seasoned experts cannot agree on the definition of cause and effect or on how one should estimate such relationships from observational studies?'

Things have changed in the past two decades. Today, experts agree (some unwittingly) on almost every aspect of causal analysis; controversies have given way to theorems, paradoxes have been resolved, and estimation problems have been algorithmitised.

James Lee is right in starting the discussion by introducing new notation. Most controversies of the past have originated with notational confusion, and most mistakes today stem from the belief that probability calculus is sufficient for handling causal relations. The insufficiency of probability (and statistics) is traumatic to most researchers in the field of data analysis because our schooling has given us the illusion that, first, a joint density function of all observed variables is the ultimate source of all knowledge and, second, everything that can be inferred from data can be inferred using the mathematical machinery of probability and statistics.

Although the *do*-operator and expressions of the type $P(mud \mid do(rain))$ may appear to be an unnecessary, if not offensive infringement on probability theory, the causal diagrams associated with such expressions convey their meaning vividly and ambiguously, especially to researchers familiar with path analysis and structural equation models (SEMs). I strongly recommend therefore that researchers invest the time in acquiring the few fundamental graphical tools necessary for causal analysis. *D*-separation is one such tool, without which one is at a loss as to what the testable implications are of a given model, whether a variable Z qualifies to serve as an instrumental variable or whether two proposed models are statistically indistinguishable. The back-door criterion is another, with the help of which one can tell immediately whether a causal effect can be estimated by adjustment on observed covariates.

However, the theory of causal diagrams differs in two fundamental aspects from conventional SEM. First, no commitment is made to linearity or to any parametric representation of the equations—these remain still qualitative. Second, the causal assumptions that go into the diagram are precisely defined and, contrary to conventional practice, are not conflated with their statistical surrogates.

DOI: 10.1002/per

In the sequel, I will summarise the tools that the new theory of causation offers to researchers: For details, please see (Pearl, 2009; 2010; 2012a).

Summary of capabilities

1. Tools for reading and explicating the causal assumptions embodied in SEM models as well as the set of assumptions that support each individual causal claim.
2. Methods of identifying the testable implications (if any) of the assumptions in (1) and ways of testing not the model in its entirety but the testable implications of the assumptions behind each causal claim.
3. Methods of deciding, prior to taking any data, what measurements ought to be taken, whether one set of measurements is as good as to another, and which measurements tend to bias our estimates of the target quantities.
4. Methods for devising critical statistical tests by which two competing theories can be distinguished.
5. Methods of deciding mathematically if the causal relationships of interest are estimable from nonexperimental data and, if not, what additional assumptions, measurements or experiments would render them estimable,
6. Methods of recognising and generating equivalent models.
7. Generalisation of SEM to categorical data and nonlinear interactions.
8. A formal solution to the problem of 'external validity' (Campbell & Stanley, 1963), that is, under what conditions can results from an empirical study be transported to another environment, differing from the first, how the results should be calibrated to account for the differences and what measurements need be taken in each of the two environments to license the transport (Pearl & Bareinboim, 2011).
9. A simple, causally based solution to the so called 'Mediation Problem', taking the form of estimable formulas for direct and indirect effects that are applicable to both continuous and categorical variables, linear as well as nonlinear interactions.

The mediation formula

This last result deserves further discussion because the problem of mediation is extremely important in personality research for it unveils the mechanisms that mediate between causes and effects.

The analysis of mediation has long been a thorny issue in the social and behavioural sciences (Baron & Kenny, 1986; MacKinnon, 2008) primarily because the distinction between causal parameters and their regressional surrogates was too often conflated (Pearl, 2012b). The difficulties were amplified in nonlinear models, where interactions between pathways further obscure their distinction.

The nonparametric analysis now permits us to define the target quantity in a way that reflects its actual usage in decision-making applications. For example, if our interest lies in the fraction of cases for which mediation was sufficient for the response, we can pose that very fraction as our target question, whereas if our interest lies in the fraction of responses for which mediation was necessary, we would pose this fraction as our target question effect (Pearl, 2001, 2012b).

In both cases, we can dispose of parametric analysis altogether and ask under what conditions can the target question be identified/estimated from observational or experimental data. One can further show that if certain conditions of 'no unmeasured confounders' hold, a simple mediation formula can be derived that captures the effects of interest. The mediation formulas are applicable to both continuous and categorical variables, and can consistently be estimated from the data.

I commend James Lee for illustrating so vividly the power of causal diagrams in a language familiar to personality researchers. I am hopeful that readers will appreciate both the transparency of the model and the power of the approach.

Does a Directed Acyclic Graph Define Causality?

ROLF STEYER

Friedrich-Schiller-Universität, Jena, Germany

rolf.steyer@uni-jena.de

Abstract: It is argued that the theory of directed acyclic graphs (DAGs) does not define causality, although a DAG—just like a structural equation model (SEM)—can be used as a causal model if appropriate assumptions are made. The causal meaning of a DAG (and a SEM) does not come from the definition of a DAG (or the SEM). Both are just statistical models not differing in this respect from the ANOVA model. The causal meaning comes from the theory of causality within which we postulate certain relationships between the variables included in the DAG or SEM and the variables not included. Neither DAGs nor SEMs address these relationships. Copyright © 2012 John Wiley & Sons, Ltd.

Reading this paper and feeling the enthusiasm for causality reminds me of the time when I learned about SEM, back in 1979. After a workshop with Karl Jöreskog, Dag Sörbom and Bengt Muthén, I felt very excited about the new possibilities of causal modelling with latent variables. This enthusiasm has been alive ever since and still motivates me working hard every day. Guess on what? Right, causality! So, welcome to the club, James!

DOI: 10.1002/per

Soon after this first experience, I began asking the following: What is the difference between an ordinary linear regression model and a causal or structural equation model? In special cases, they look identical: linear function of the regressor; zero correlation between regressor and error term; and zero expectation of the error term. Asking all experts and reading all books and papers on the topic, I found many answers. Let me mention three of them. First, according to Wold, the relationship described in a SEM 'is then defined as causal if it is theoretically permissible to regard the variables as involved in a fictive controlled experiment' (Wold, 1954, p. 166). Second, Goldberger wrote that the parameters in a SEM are 'the fundamental parameters of the mechanism that generated the data' (Goldberger, 1973, p. 3). Third, many other econometricians argued that the error term is not a residual but *a disturbance*, consisting of all other influences on the corresponding regressand not included in the model. Pearl's definition of a causal effect (Pearl, 2009, p. 70) rests on exactly this third answer.

Although each of the three answers has an element of truth, I have never been happy with them. In the ideal case, the randomised experiment guarantees that the effect $E(Y|X=1) - E(Y|X=0)$ of a dichotomous treatment variable X (with values 0 and 1) on the outcome variable Y is the *total causal treatment effect*. But how does this help in identifying a causal effect in a quasi-experiment in which, by definition, a unit is not randomly assigned to one of the treatment conditions? And how does it help in identifying *direct treatment effects*? (In the meantime, we know that randomisation does *not guarantee* that $E(Y|X=1, M=m) - E(Y|X=0, M=m)$ is the $(M=m)$-conditional direct treatment effect, which questions all those applications of the Baron–Kenny approach to mediation (Baron & Kenny, 1986), that only use the treatment, the mediator(s) and the outcome variable in the SEM (see, e. g., Mayer, Thömmes, Rose, Steyer, & West, 2012). In fact, the problem is the collider problem described by Lee.) The second answer is just a metaphor that helps if we generate data in a simulation. But what does it imply in real-life studies? Finally, the third answer is contradictory. Pearl's own words are as follows: 'These disturbance terms represent independent background factors that the investigator chooses not to include in the analysis' (Pearl, 2009, p. 68). What is it that I do not buy?

According to Pearl, 'each child–parent family in a DAG G represents a deterministic function $x_i = f_i\ (pa_i\ ,\varepsilon_i\ )$, ...' Hence, together with the pa_i, the disturbance ε_i *deterministically* determines x_i. The problem is that the deterministic equation $x_i = f_i\ (pa_i\ ,\varepsilon_i\ )$ can be true only if the ε_i also 'represent' (whatever this term may mean) all the mediators transmitting the effects of the pa_i to the x_i. Otherwise, $x_i = f_i\ (pa_i\ ,\varepsilon_i\ )$ cannot be true. However, mediators will correlate with the pa_i. (Just think of a perfect causal chain.) Hence, ε_i and pa_i cannot be independent.

Because Pearl's $do(x)$ terms and his definition of a causal effect rest on the disturbance term, both are contradictory as well. This is why I do not buy the claim that a DAG defines causality, although they are an important tool in the analysis of causal effects. Exactly the same applies to SEMs and to the general linear model (GLM). As statistical models, all of them are important tools and have the potential to describe causal dependencies. However, they are no theories of causality.

To repair this defect, Pearl's 'independent background factors' need an explicit place in a theory of causal effects. In the real world, they all too often act as confounders, but up to date, they do not have an adequate representation in the theory of DAGs. Therefore, they spook around as unsettled ghosts poisoning the discussions on statistics and causality (see, e. g. Bollen & Pearl, in press; Pearl, 2009, p. 104).

Suppose we are interested in the effect of a treatment variable X on an outcome variable Y and we assume stochastic independence between X and all *pretreatment variables*, then it is this assumption that allows us to give a causal interpretation of the regression of Y on X and of the conditional distribution of Y on X. There is no need to include all pretreatment variables in a DAG; they are too many, anyway. A DAG and the random variables constituting it have to be embedded in a *probability space* that represents the random experiment to be discussed. The pretreatment variables (potential confounders) refer to the same probability space and so do omitted mediators. If we endow the probability space with a *filtration* (a nondecreasing family of σ-algebras), a fundamental concept in the theory of stochastic processes (see, e. g. Klenke, 2008), then we have the mathematical prerequisites to *order* all our variables so that the term 'pretreatment variable' has a formal representation (see, e. g. Steyer, 1984, 1992) not resting on a concept of causality (which would be circular; this is the problem if we use a DAG for such an ordering). Also, our distinction between *covariates* and *intermediate variables* or *possible confounders* and *possible mediators* will a have a mathematical foundation if based on the concept of a filtration.

Applying a DAG requires specifying a complete causal model for all variables involved. Only then can we decide which variables may be omitted (see Lee's discussion on colliders) and identify the causal effect of a variable X on a variable Y. An important implication of the stochastic theory of causal effects outlined previously (see, e.g. Steyer, Mayer, & Fiege, in press and Steyer, Fiege, & Mayer, in press) is that it is applicable without complete knowledge about the causal relations between variables that act in the random experiment considered. A gross ordering such as 'prior to treatment', 'in between treatment and outcome' and 'posterior to treatment' is often sufficient.

Let me add a few remarks indicating where I disagree with Lee. If appropriately constructed, latent variables *can* be considered causes of its manifest measures. This point of view is also shared by Bollen and Pearl (in press), and it can be proven within the theory of stochastic causality outlined previously. (Actually, I have much more problems with causation among latent variables, at least if there is not a clear time sequence between them, such as in a longitudinal study.) In contrast to Lee, Bollen and Pearl (2012) also do *not claim* that SEMs are inadequately formalised. Instead, they treat a SEM as a special sophisticated DAG, and I can follow them in this respect. For the reasons mentioned previously, I do not think that the graphical

DOI: 10.1002/per

framework 'captures human intuitions about causality in the form of consistent mathematical axioms'. ('Background factors' is not a mathematically well-defined term.) However, if embedded in a stochastic theory of causality, DAGs are among the greatest achievements in the last 50 years. They can be very useful, as can SEM, the GLM and hierarchical linear models.

A final caveat is as follows: neither DAGs nor SEMs or any other statistical model are techniques for discerning causation in observational data. They only help to derive logical implications of a model and test them, to some degree, against reality (see myth # 1 in Bollen & Pearl, in press). For the rest, let me congratulate the author for this great and impressive paper.

Welcome Clarity for Muddy Waters

ALEXANDER WEISS AND TIMOTHY C. BATES

School of Philosophy, Psychology and Language Sciences, Department of Psychology, University of Edinburgh, UK

alex.weiss@ed.ac.uk

Abstract: In his article, Lee advocates using directed acyclic graphs to test causal theory in differential psychology. We applaud Lee's efforts to introduce graph theory and causal hypotheses to differential psychology. We agree that these methods lead to better understanding of the mechanisms' underlying behaviour. We thus join him in advocating the use of these methods that, although require more creative effort, honestly confront and overcome the problems of assigning causality that can plague differential psychology and public policy research. Copyright © 2012 John Wiley & Sons, Ltd.

Testing causal hypotheses against observational data is central to progress in differential psychology and the social sciences more generally. As noted by Lee, the lack of causal progress in the social sciences flows not from a lack of statistical tools but rather from which tools are used and how. Lee's most important message is that whereas social scientists are taught that 'correlation does not imply causation', they are not taught what correlation **does** imply, namely an unresolved causal structure (Shipley, 2000). Lee lays out the requirements for causal research, namely how causal hypotheses must be expressed as directed acyclic graphs (DAGs) or equivalent structures, and how theories thus expressed can then be objectively compared.

Since works such as Kerlinger's (1964) 'Foundations of Behavioral Research', students have been trained to 'control' variables. This method has become embedded in the paradigm of psychological science along with the assumption that control leads to conclusions. In fact, however, it has long been recognised that inappropriate controls mischaracterise cause and effect (e.g. Meehl, 1992). In particular, apparently innocuous control of a variable that is influenced by traits under study can induce false associations between these traits of interest, which, if interpreted as real, can have harmful consequences (Bingham, Heywood, & White, 1991; Figueredo, Hetherington, & Sechrest, 1992). Beyond this, statistical control cannot, even in principle, test causal assumptions (Pearl, 2000). Knowledge about just how critically limited epidemiological models are in identifying causes is far from widespread, and hopefully, Lee's article will spread this knowledge much further. The calculus of causal theory, in proving the consequences of lack of control and inappropriate statistical control, as well as the solutions to these problems, places causal theorising on a firm mathematical and logical footing. There simply can no longer be any place for theories not expressed, whether in words or figures, as DAGs making these effects explicit.

Confidence in the progress of theory

One important consequence of a logical framework for contrasting causal hypotheses is that the crisis of confidence that found its expression in the postmodern proposition that as social constructions, all theories have equivalent value is set aside. Although scientists still cannot know if they have found the truth simply by dint of applying DAGs, they can determine which of two competing models is closer to that truth. If a mechanism exists to objectively and iteratively select causal models that are not simply different, but which are more complete in an objective sense, this may be the biggest impact of all the changes that Lee's paper lays out.

Theory generation

If statistical tools allow us to test causal hypotheses, they also highlight the requirement for researchers to generate theories. Importantly, when a model finds itself containing a correlation, as with rain and mud, one must 'do(mud)' and measure the effects or lack thereof of this treatment on the likelihood of rain. We hope that this form of expression spreads widely and that readers come to expect articles to be expressed in this fashion, compelling researchers to make clear the causal process they are predicting, be it do(school) or do(genetic polymorphism) or do(neuroticism). But what if they do not? One adverse consequence of not translating causal theory into practice is that of targeting outcomes rather than processes leading to outcomes. For instance, targeting school grades instead of effective teaching can have perverse consequences as factors not included in causal models come into play. These factors may include teachers invalidating tests as indicators of knowledge by teaching to the test or, worse, purchasing the answers to exams (Vasagar, 2011). Lest other fields feel smug and secure in their methods, such errors remain common in

DOI: 10.1002/per

areas such as medicine that are more used to thinking causally. Variables merely associated with a disease may become proposed targets for intervention, sometimes to humorous effect (Cohen et al., 2000). Such problems should be expected in the study of any system that has multiple causes (see the First Law of Ecology in Hardin, 1963).

What are common factors?

We will finish with a discussion of Lee's statement regarding common factors. Lee 'allows a factor to play the role of cause or effect in graphs depicting the relations among high-level emergent entities.' The meaning ascribed to a common factor such as extraversion is basic to psychology, and we appreciated the nuanced claim that a common factor may 'play the role' of a cause. Lee is, here as elsewhere, taking causal reasoning seriously. Latent variable on structural equations modelling diagrams represent emergent properties of their indicators, but, like correlations, they also represent as-yet unresolved causal structures and must be explained by mechanisms. In psychology, explanations of constructs such as in-group favouritism are often given in what in Lee's terms would be long-form labels for the emergent property, or even as additional indicators. Lee's conceptualisation thus refocuses our attention on the need to hypothesise causes for latent variables, not simply generate labels or additional indicators for them. Just as in physics, the emergent properties of water are accounted for by non-wet, nonliquid causes, so too personality domains such as extraversion must be accounted for by layers of mechanisms, from biology through typical characteristics to the objective biography of behaviour (McCrae, 1996), rendering a set of objectively specified and parameterised mechanisms generating behaviour on the fly as these systems are run in real environments (Lewis & Bates, 2011).

Causal hypothesising and testing is, of course, no 'royal road' to knowledge: while the means of testing causal mechanisms are established, causal hypotheses cannot be generated automatically. As Gödel (1962) demonstrated, steps towards completeness require creative mental effort that is not automatable. The power of modern differential psychology, then, depends on specifying theory in directed acyclic graphs permitting causal inference, and we commend Lee's article to as broad a readership as possible.

AUTHOR'S RESPONSE

Causes and Effects of Common Factors

JAMES J. LEE[1,2*]

[1]*Vision Lab, Department of Psychology, Harvard University, USA*
[2]*Cognitive Genomics Lab, BGI-Shenzhen, China*

son.of.jorel.34@gmail.com

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

Key words: personality; causality; directed acyclic graph; structural equation modelling; behavioural genetics

INTRODUCTION

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 towards the praise garnered by my modest contribution.

*Correspondence to: James J. Lee, Vision Lab, Department of Psychology, Harvard University, Cambridge, MA, USA.
E-mail: son.of.jorel.34@gmail.com

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 that can be regarded as a corollary of what Pearl calls the *back-door theorem*. Unfortunately, the statement of the theorem by **Kievit, Waldorp, Kan, and Wicherts** contains several errors and ambiguities. Their hypothesis (i) states that all direct causes of X must be

DOI: 10.1002/per

statistically controlled 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, because 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—i.e., 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 nondirected paths between *X* and *Y*, then the expected value of *Y* upon setting *X* equal to x_1 is

$$E(Y|do(x_1)) = \iint yp(y|x_1, z)p(z)dzdy$$

where $p(.)$ is the appropriate probability density. The form of this expression tells us that we must average over all possible values of *Z*; picking just one value may lead to a "stratum-specific" estimate. 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. This observation leads 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" 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 causal 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's concern over the number of partial correlations (conditional independencies) confuses the testing of a *priori* causal models with a model-free evaluation of all possible partial correlations. For various reasons, I will not address the latter approach.

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 10, not 20, 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 attacks the notion of a causal effect of *X* on *Y* as reflecting the sensitivity of *Y* to the randomization of X, 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)

Steyer's more specific point regarding indirect or conditional causal effects misses the mark for the same reason. For instance, in a randomized experiment examining the effect of sleep deprivation on state anxiety, 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)**. A do operation on the appropriate DAG, however, 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)**.

Steyer (along with **Jackson and Spain**) accuses 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 flavour.

DOI: 10.1002/per

Steyer (1984) claims to offer an alternative account of causality based on probability theory. Unfortunately, I find this account to be quite obscure. For example, Steyer's condition 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. One of my statistics instructors once defined his subject as the use of finite data to infer the parameters of probability distributions; taking this notion seriously, suppose that the sample size is so large that we can estimate the parameters of the relevant probability distribution with perfect accuracy. 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 towards 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 nonstatistical: *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 sceptical of attempts to build a formal account of causality on purely probabilistic grounds.

The foundational importance and distinctiveness of causal notions imply 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. Pearl's (2009) treatments of instruments and propensity scores may be the most lucid that I have seen anywhere in the literature. Although I am more cautious than **Davey Smith**, I am on the whole sympathetic towards his use of genetic variants as instruments and foresee the fruitfulness of this approach in the investigation of whether the biological correlates of a certain personality trait are indeed causes of that trait.

I share the reservations of **Kievit, Borsboom, Jackson, and Spain** and **Davey Smith** regarding cross-sectional studies of high-level variables that do not incorporate some special feature (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 blockacyclic model by sufficiently fine-grained distinctions among time points within individuals. The importance of causal notions does not diminish, however, 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 time points. Note that each directed edge in Figure 5 of Cramer et al. (this issue) represents temporal order rather than a cause–effect relation. 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 validity of the premises is in doubt, then we cannot draw any firm conclusions. Such failures, however, are no reason to denigrate *logic*. The logic of causality is no different in this respect. Likewise, if *any* conclusions are warranted at all, their warrant must rest on both the requisite premises and the appropriate governing logic. Causal conclusions are again no different.

For instance, how do we know that smoking causes death? **Davey Smith** cites some new evidence on this point, but let us examine a more established element 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 the expected value

$$E_G(death, smoking|G = g) = E_G(death, smoking|see(G = g)),$$

but wish to infer

$$E_G(death, smoking|do(G = g)).$$

This substitution is legitimate because it was nature that *fixed* the genome to be identical for both members of a twinship. That is, because the same chain of events produced the genotype of each twin, the nodes in the causally prior subgraph

DOI: 10.1002/per

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, the rest of us 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 0.500. This success rate should make GWAS the envy of other biomedical and behavioural scientists who must deal with observational data. We are thus entitled to ask: what special features of GWAS make causal inference, if not infallible, at least feasible? 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, arguing for its validity and demonstrating the identification of the desired quantity are essential to the justification of any empirical study claiming to advance our causal knowledge.

This is why it is 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**), when in fact *there is no non-graphical alternative that avoids these challenges*. As Pearl (2009, pp. 173–200) showed 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.

> 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 Antonin Scalia has the intellect to be a leader of the 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 factor analysis is an attempt to quantify folk-psychological attributes of this kind, others continue to question the usefulness of this enterprise (**Borsboom**).

It 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 for now, the evaluation of "construct validity" remains one of the case-by-case, non-automatable aspects of behavioural 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 behavioural 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 levelled this charge of aggregating an atheoretical hodgepodge. Because 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 gene-trait mapping studies 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 (Meuwissen & Goddard, 2010), 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?

DOI: 10.1002/per

Although this thought experiment regarding biological significance does not seem to illustrate any algorithmic principle of construct validation, it suggests that uses of 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 the entire literature on construct validity as the demonstration of correlations with external variables that common sense already tells us to be correlated with the target attribute. The kinds of biological insights that might follow from genetic research, however, are certainly no matters of common sense.

THE PROBLEM OF SELECTION BIAS

I agree with **Aspendorf** that selection bias is a taboo subject among behavioural scientists. It was perhaps 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".

Jackson and Spain do not distinguish selection bias from confounding. Figure 1 hopefully clarifies the difference. Refraining from any assumptions about the graph containing *X* and *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* would suffice to identify any $X \rightarrow Y$ causal effect in the absence of selection bias. If selection bias is in fact present, more assumptions are necessary regarding the *d*-separation of all paths between *X* and *Y* passing through study appearance. Bareinboim and Pearl (in press) studied 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 remainder may not be enough. This difficulty admittedly complicates studies of ageing.

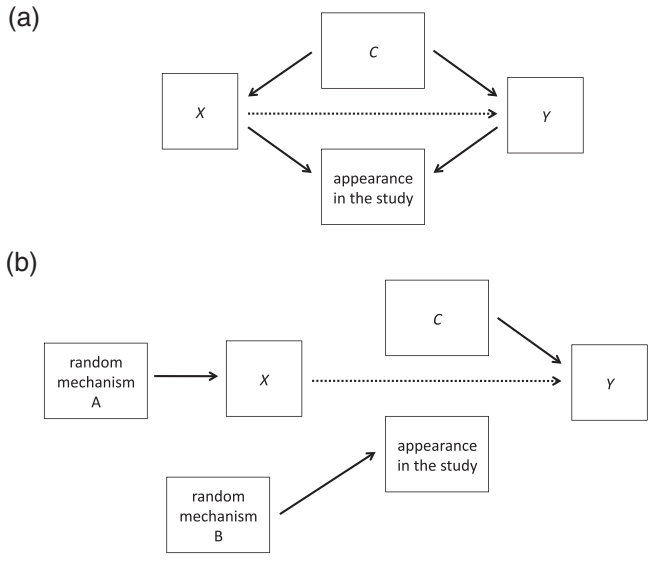

Figure 1. DAGs representing the distinction between confounding and selection bias. (a) The case of no *a priori* assumptions regarding the relation between *X* and *Y*. (b) Randomization of *X* deletes the $C \rightarrow X$ edge. Randomization of study appearance deletes both $X \rightarrow$*appearance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

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 encompass transportability across environments (countries differing in climate and infrastructure) and nonlinearity (causal effects depending on "where we are"), perhaps the target article should have discussed these matters in more detail. 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. 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 the contrary. Scalia's increased g may allow him to rationalize positions that he holds for nonintellectual reasons and find the holes in 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 law-like 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.

DOI: 10.1002/per

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 22—one trait-affecting locus per autosome—the number of possible genotypes already exceeds the current size of the human population by more than fourfold. 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 harbour 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 non-linear 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.

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 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 programme. In recent years, we have perhaps neglected Allport's (1937) vision of an allencompassing 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 have been used to follow up mapping studies, tracing the intermediate mechanisms between genotype and phenotype (Pomerantz et al., 2009;Musunura 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". The ease with which depicted 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 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 marvelled 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 graphical taxonomy of correlations, which states that a non-coincidental correlation between X (phenotype) and Y (environment) must be attributable to

1 X causing Y,
2 Y causing X,
3 X and Y being effects of a common cause or
4 X and Y being causes of a common effect that has been statistically controlled.

To explain biological adaptation, pre-Darwinians often invoked explanation (3), ascribing the role of common cause to a benevolent Creator. After Darwin and Mendel, many biologists favoured explanation (2), invoking hypothetical mechanisms by which environmental circumstances might mould the causes of phenotypic variation. Weismann (1893) gave reasons for rejecting Lamarckianism and other explanations within this class that remain cogent today. Explanation (1), in which organisms seek out or create environments promoting their own fitness, obviously cannot suffice because such niche-seeking capacities are themselves complex

DOI: 10.1002/per

adaptations. It was the genius of Darwin to realize the power of explanation (4): 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 phenotypic consequences 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. **Davey Smith** and I are confident that where nature has succeeded, patient and ingenious human scientists will be able to follow.

REFERENCES

Ackerman, P. L., & Heggestad, E. D. (1997). Intelligence, personality, and interests: Evidence for overlapping traits. *Psychological Bulletin*, *121*, 219–245.

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

Almlund, M., Duckworth, A., Heckman, J., & Kautz, T. (2011). Personality psychology and economics. In: E. A. Hanushek, S. Machin, & L. Woessmann (Eds), *Handbook of the economics of education*. Amsterdam: Elsevier.

Antonakis, J., Bendahan, S., Jacquart, P., & Lalive, R. (2010). On making causal claims: A review and recommendations. *The Leadership Quarterly*, *21*(6), 1086–1120. DOI: 10.1016/j.leaqua.2010.10.010

Baron, R., & Kenny, D. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*, 1173–1182.

Bingham, R. D., Heywood, J. S., & White, S. B. (1991). Evaluating schools and teachers based on student performance: Testing an alternative methodology. *Evaluation Review*, *15*, 191–218.

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

Bodmer, W. (2003). R. A. Fisher, statistician and geneticist extraordinary: a personal view. *International Journal of Epidemiology*, *32*, 938–942.

Bollen, K. A., & Pearl, J. (in press). Eight myths about causality and structural equation models. In S. Morgan (Ed.), *Handbook of Causal Analysis for Social Research*. New York: Springer.

Box, J. F. (2010). Commentary: On R. A. Fisher's Bateson lecture on statistical methods in genetics. *International Journal of Epidemiology*, *39*, 335–339

Campbell, D., & Stanley, J. (1963). *Experimental and quasi-experimental designs for research*. Chicago: Wadsworth Publishing.

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.

Cohen, D., Spear, S., Scribner, R., Kissinger, P., Mason, K., & Wildgen, J. (2000). 'Broken windows' and the risk of gonorrhea. *American Journal of Public Health*, *90*, 230–236.

Costa, P. T., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) Professional Manual*. Odessa, FL: Psychological Assessment Resources.

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist*, *12(*11*)*, 671–684.

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

Dahlhaus, R., & Eichler, M. (2003). Causality and graphical models for time series. In: P. Green, N. Hjort, & S. Richardson (Eds), *Highly structured stochastic systems* (pp. 115–137). Oxford: University Press.

Davey Smith, G. (2006). Capitalising on Mendelian randomization to assess the effects of treatments. *James Lind Library Bulletin: Commentaries on the history of treatment evaluation*. (www.jameslindlibrary.org).

Davey Smith, G. (2010). Mendelian randomization for strengthening causal inference in observational studies: Application to gene by environment interaction. *Perspectives on Psychological Science*, *5*, 527–545.

Davey Smith, G. (2011a). Epidemiology, epigenetics and the 'gloomy prospect': Embracing randomness in population health research and practice. *International Journal of Epidemiology*, *40*, 537–562.

Davey Smith, G. (2011b). Random allocation in observational data: How small but robust effects could facilitate hypothesis-free causal inference. *Epidemiology*, *22*, 460–463.

Davey Smith, G., & Ebrahim, S. (2002). Data dredging, bias, or confounding (editorial). *British Medical Journal*, *325*, 1437–1438.

Davey Smith, G., & Ebrahim, S. (2003). 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal of Epidemiology*, *32*, 1–22.

Davey Smith, G., Lawlor, D. A., Harbord, R., Timpson, N. J., Day, I., & Ebrahim, S. (2008). Clustered environments and randomized genes: A fundamental distinction between conventional and genetic epidemiology. *PLoS Medicine*, *4*, 1985–1992.

Dawid, A. P. (2008). Beware of the DAG. *Journal of Machine Learning Research: Workshop and Conference Proceedings*, *6*, 59–86.

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

DeYoung, C. G., Quilty, L. C., & Peterson, J. B. (2007). Between facets and domains: 10 aspects of the Big Five. *Journal of Personality and Social Psychology*, *93(*5*)*, 880–896.

Draganski, B., Gaser, C., Busch, V., Schuierer, G., Bogdahn, U., & May, A. (2004). Neuroplasticity: Changes in gray matter induced by training. *Nature*, *427*, 311–312.

Eichler, M. (2007). Granger-causality and path diagrams for multivariate time series. *Journal of Econometrics*, *137*, 334–353.

Ellis, J. L., & Junker, B. W. (1997). Tail-measurability in monotone latent variable models. *Psychometrika*, *62*: 495–523.

Figueredo, A. J., Hetherington, J., & Sechrest, L. (1992). Water under the bridge: A response to Bingham, Heywood, and White. *Evaluation Review*, *16*, 40–62.

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. Reprinted in *International Journal of Epidemiology*, *39*, 329–335.

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

Foster, E. M. (2010). Causal influence and developmental psychology. *Developmental Psychology*, *46*, 1454–1480.

Freathy, R., Kazeem, G., Morris, R., Johnson, P., Paternoster, L., Ebrahim, S., … Munafò, M. (2011). Genetic variation at CHRNA5-CHRNA3-CHRNB4 interacts with smoking status to influence BMI. *International Journal of Epidemiology*, *40*, 1617–1628.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York: Cambridge 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.

Gödel, K. (1962). *On formally undecidable propositions of principia mathematica and related systems* (First English edn). Edinburgh, UK: Oliver and Boyd.

Goldberger, A. S. (1973). Structural equation models: An overview. In A. S. Goldberger, & O. D. Duncan (Eds), *Structural equation models in the social sciences*, (pp. 1–18). New York, NY: Seminar Press.

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

Hamaker, E. L., Nesselroade, J. R., & Molenaar, P. C. M. (2007). The integrated trait-state model. *Journal of Research in Personality*, *41*, 295–315.

Hardin, G. (1963). The cybernetics of competition: A biologist's view of society. *Perspectives in Biology and Medicine*, *7*, 58–84.

Heckman, J. J. (2005). The scientific model of causality. *Sociological Methodology*, (*35*), 1–98.

Hindorff, L. A., MacArthur, J., Wise, A., Junkins H. A., Hall, P. N., Klemm, A. K., & Manolio TA. (2012). *A catalog of published genome-wide association studies*. Available at: www.genome.gov/gwastudies, retrieved 10 April 2012.

Holland, P. (1986). Statistics and causal inference. *Journal of the American Statistical Association 81*, 945–970.

Jackson, J. J., Thoemmes, F., Jonkmann, K., Lüdtke, O., & Trautwien, U. (2012). Military training and personality trait development: Does the military make the man or does the man make the military? *Psychological Science*, *23*, 270–277.

Johnson, W. (2007). Genetic and environmental influences on behavior: Capturing all the interplay. *Psychological Review*, *114(2)*, 423–440.

Johnson, W., Penke, L., & Spinath, F. M. (2011). Heritability in the era of molecular genetics: Some thoughts for understanding genetic influences on behavior. *European Journal of Personality Special Issue Target Article*, *25*, 255–266.

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, F., Graefen, A., Ball, M., Matzas, M., Boisguerin, V., Mainxer, F., … Zink, A. (2012). New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing. *Nature Communications*, *3*, 698.

Kendler K. S., Gardner C. O. (2010). Dependent stressful life events and prior depressive episodes in the prediction of major depression. *Archives of General Psychiatry 67*, 1120–1127.

Kennedy, P. (1985). *A guide to econometrics* (2nd edn). Oxford: Basil Blackwell.

Kerlinger, F. N. (1964). *Foundations of behavioral research*. New York, NY: Holt, Rinehart and Winston, Inc.

Klenke, A. (2008). *Probability theory—A comprehensive course*. London: Springer.

Lauritzen, S. L. (1996). *Graphical models*. Oxford: Oxford University Press.

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

Lewis, G. J., & Bates, T. C. (2011). From left to right: How the personality system allows basic traits to influence politics via characteristic moral adaptations. *British Journal of Psychology*, *102*, 546–558.

Lipton, P. (2004). *Inference to the best explanation*, 2nd edn. Abingdon, UK: Routledge, 2004.

MacKinnon, D. (2008). *An introduction to statistical mediation analysis*. New York: Lawrence Erlbaum Associates.

Markus, K. A., & Borsboom, D. (2011). Reflective measurement models, behavior domains, and common causes. *New Ideas in Psychology*. DOI: 10.1016/j.newideapsych.2011.02.008.

Mayer, A., Thömmes, F., Rose, N. Steyer, R., & West, S. G. (2012). *Theory and analysis of total, direct and indirect causal effects*. Manuscript submitted.

McCrae, R. R. (1996). Integrating the levels of personality. *Psychological Inquiry*, *7*, 353–356.

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

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.

Meehl, P. E. (1992). A funny thing happened to us on the way to the latent entities. In E. I. Megargee, & C. D. Spielberger (Eds.), *Personality assessment in America: A retrospective on the occasion of the fiftieth anniversary of the Society for Personality Assessment* (pp. 113–125). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

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 Perspectives*, *2*, 201–218.

Molenaar, P.C.M., & Campbell, C.G. (2009). The new person-specific paradigm in psychology. *Current Directions in Psychology*, *18*, 112–117.

Motti-Stefanidi, F., Asendorpf, J. B., & Masten, A. S. (2012). The adaptation and well-being of adolescent immigrants in Greek schools: A multilevel, longitudinal study of risks and resources. *Development and Psychopathology*, *24*, 451–473.

Munafo, M. R., Timofeeva, M. N., Morris, R. W., Prieto-Merino, D., Sattar, N., Brennan, P., … Davey Smith, G. (2012). Association between genetic variants on chromosome 15q25 locus and objective measures of tobacco exposure. *Journal of the National Cancer Institute*. DOI: 10.1093/jnci/djs191.

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

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge, England: Cambridge University Press.

Pearl, J. (2001). *Direct and indirect effects*. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence* (pp. 411–420). San Francisco, CA: Morgan Kaufmann, San Francisco.

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

Pearl, J. (2010). The foundations of causal inference. *Sociological Methodology*, *40*, 75–149.

Pearl, J., & Bareinboim, E. (2011). Transportability across studies: A formal approach. *Proceedings of the 25th AAAI Conference on Artificial Intelligence*, (pp. 247–254). Menlo Park, CA: AAAI Press.

Pearl, J. (2012a). The causal foundations of structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (68–91). New York: Guilford Press.

Pearl, J. (2012b). The causal mediation formula—A guide to the assessment of pathways and mechanisms. Online, *Prevention Science*, DOI: 10.1007/s11121-011-0270-1, March 2012.

Penke, L., Borsboom, D., Johnson, W., Kievit, R. A., Ploeger, A., & Wicherts, J. M. (2011). Evolutionary psychology and intelligence research cannot be integrated the way Kanazawa (2010) suggests. *American Psychologist*, *66*, 916–917.

Pomerantz, M. M., Ahmadiyeh, N., Jia, L., Herman, P., Verzi, M. P., Doddapaneni, H., ... Freedman, M. (2009). The 8q24 cancer risk variant rs6983267 shows long-range interaction with *MYC* in colorectal cancer. *Nature Genetics*, *41*, 882–884.

Provine, W. B. (1986). *Sewall Wright and evolutionary biology*. Chicago: University of Chicago Press Books.

Rasmussen, M., Li, Y., Lindgreen, S., Pedersen, J. S., Albrechtsen, A., Moltke, I., ... Willerslev, E. (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., ... Pääbo, S. (2010). Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature*, *468*, 1053–1060.

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(4)*, 313.

Rubin D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, *100*, 322–331.

Rutter M. (2007). Proceeding from observed correlation to causal inference. *Perspectives on Psychological Science*, *2*, 377–395.

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

Sheehan, N. A., Didelez, V., Burton, P. R., & Tobin, M. D. (2008). Mendelian Randomisation and causal inference in observational epidemiology. *PLoS Medicine*, *5*, e177.

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.

Smilie, L. D., Cooper, A., Wilt, J., & Revelle, W. (in press). Do extraverts get more bang for the buck? Refining the affective-reactivity hypothesis of extraversion. *Journal of Personality and Social Psychology*.

Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, prediction, and search*. Cambridge, MA: MIT Press.

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.

Steyer, R. (1992). *Theorie Kausaler Regressionsmodelle [Theory of causal regression models]*. Stuttgart, Germany: Fischer.

Steyer, R., Fiege, C., & Mayer, A. (in press-a). Causal inference. In: A. C. Michalos (Ed.), *Encyclopedia of quality of life research*. Heidelberg: Springer.

Steyer, R., Mayer, A., & Fiege, C. (in press-b). Total, direct and indirect effects. In: A. C. Michalos (Ed.), *Encyclopedia of quality of life research*. Heidelberg: Springer.

Thoemmes, F., & Kim, E. (2011). A systematic review of propensity score methods in the social sciences. *Multivariate Behavioral Research*, *46*, 90–118. DOI: 10.1080/00273171.2011.540475

Timpson, N. J., Wade, K. H., & Davey Smith, G. (2012). Mendelian randomization: Application to cardiovascular disease. *Current Hypertension Reports*, *14*, 29–37.

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.

Vasagar, J. (2011, Thursday 8 December 2011). *Exam boards scandal: the economic pressures that broke the system*, *The Guardian*. Retrieved from http://www.guardian.co.uk/education/2011/dec/08/exam-boards-scandal-analysis.

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.

Wang, Y., Broderick, P., Matakidou, A., Eisen, T., & Houlston, R. S. (2011). Chromosome 15q25 (*CHRNA3-CHRNA5*) variation impacts indirectly on lung cancer risk. *PLoS ONE*, *6*, e19085.

Wellcome Trust Case Control Consortium. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature 447*, 661–78.

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

White, J.A. (1990). Ideas about causation in philosophy and psychology. *Psychological Bulletin*, *108*, 3–18.

Wold, H. O. A. (1954). Causality and econometrics. *Econometrica*, *22*, 162–177.

Wright, S. (1917). On the probable error of Mendelian class frequencies. *The American Naturalist*, *51*, 373–375.

Wright, S. (1920). The relative importance of heredity and environment in determining the piebald pattern of guinea pigs. *Proceedings of the National Academy of Sciences*, *6*, 320–32.

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research*, *20*, 557–585.

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.

Zhu, J., Wiener, M. C., Zhang, C., Fridman, A., Minch, E., Lum, P. Y., ... Schadt, E. E. (2007). Increasing the power to detect causal associations by combining genotypic and expression data. *PLoS Computational Biology*, *3*, e69.

DOI: 10.1002/per