Article

Theory & Psychology

Theory & Psychology
1–16

DOI: 10.1177/0959354315569832
tap.sagepub.com
SAGE

If intelligence is a cause, it is a within-subjects cause

Naftali Weinberger
University of Wisconsin

Abstract

Borsboom, Mellenbergh, and van Heerden argue that latent variables such as intelligence should be given a between-subjects causal interpretation, but not a within-subjects causal interpretation. That is, while intelligence is a cause of one subject's doing better than another on an IQ test, there is no non-comparative sense in which intelligence – as standardly measured – is a cause of an individual's performance. Here I expand upon Pearl's discussion of Simpson's paradox to show that there cannot be a cause in a population that is not a cause in at least one of its members and that, consequently, causal variables cannot have an exclusively between-subjects interpretation. The illusion that they can results from not properly distinguishing between causal and non-causal models.

Keywords

causal modeling, causation, interindividual variations, Simpson's paradox, transportability

In many psychological studies of intelligence, the researcher collects data on each participant at a single point in time and produces a model reflecting the differences in intelligence among those individuals. Do such models adequately represent the way that an individual's intelligence varies over time? Peter Molenaar (2004) has compellingly argued that models in which a latent variable such as intelligence varies among subjects in a population represent the variation of latent variables within individuals only given assumptions that are not typically justified.[1] It would be fallacious, for example, to infer that because intelligence varies among individuals in a population, there must be variation in the intelligence of at least one member of that population.

Intelligence—like other traits considered in psychometrics—is not measured directly, but is posited as a latent variable to account for the correlations among a subject's results

Corresponding author:
Naftali Weinberger, University of Wisconsin, 5185 Helen C White, 600 North Park St, Madison, WI 54706, USA.
Email: naftali.weinberger@gmail.com

on various tests. The variables measuring a subject's performance on different tests are treated as *indicators* of her value on the latent variable. Given that intelligence is not directly observed, there is room for debate regarding whether we should think of it as a real entity or as a useful fiction for making predictions. In "The Theoretical Status of Latent Variables," Borsboom, Mellenbergh, and van Heerden (2003) argue that one can make sense of psychological practice only if one understands latent variable models as making assertions about the existence of genuine causal factors.

Given Borsboom et al.'s thesis that intelligence models in psychology are committed to intelligence being a cause and Molenaar's thesis that variation between individuals is not a reliable guide to within-individual variation, it is tempting to argue as follows. Since there can be inter-individual variation in intelligence that does not correspond to variation within any individual and our causal inferences rely on probability distributions reflecting different types of variation, it is similarly possible for there to be a causal model in which intelligence is a cause of variation between subjects in a population, but not a cause acting within any individual. This possibility underlies Borsboom et al.'s suggestion that we should think of intelligence as a *between-subjects* cause explaining individual differences in their indicators rather than a *within-subjects* cause, reflecting the mechanisms at work within any particular individual.

The inference from Molenaar's (2004) thesis to the conclusion that it is possible for there to be a between-subjects cause that is not also a within-subjects cause is fallacious. Although the patterns of variation in a population need not reflect the variation within any individual, it is impossible for there to be a cause in a population that is not a cause within at least *some* of its members. This is not to deny that it is difficult to predict the effect of a causal variable in a *particular* member of a population when one does not know about that individual's properties. Even if exercise reduces the risk of heart disease in a population, it may increase the risk of heart disease in some particular member of that population. But Borsboom et al.'s thesis entails the far stronger claim that it is possible for a cause to obtain in a population as a whole, but in *none* of its members. This, I will argue, only appears possible if one does not properly distinguish between causal and non-causal relationships. While there are definitely risks in extrapolating from a population to an individual, the possibility of a between-subjects causal variable with no within-subjects interpretation is not one of them.

This article is organized as follows. I present Molenaar's results regarding the conditions under which intraindividual variation differs from interindividual variation in the first section. In the next section, I prove that there cannot be a between-subjects cause that is not also a within-subjects cause. Next, I examine Borsboom et al.'s argument that intelligence is a variable with an unproblematic causal interpretation and then I consider their reasons for thinking that intelligence is not a within-subjects cause. In the final section, I consider the implications of the previous discussion for research in psychology.

Before proceeding, a comment about the scope of this paper is in order. Although I focus on the variable of intelligence, the aim of the paper is to show that *no* causal variable can have an exclusively between-subjects interpretation. While the paper evaluates Borsboom et al.'s arguments regarding the status of intelligence, it does not further engage with the many debates pertaining to the concept of intelligence and the way it has historically been measured.

Molenaar's manifesto

In his influential article "A Manifesto on Psychology as Idiographic Science: Bringing the Person Back into Scientific Psychology, this Time Forever," Peter Molenaar (2004) argues that standard psychometric models are inadequate for predicting how individuals will develop over time, since they only measure the variation among the individuals in a population at a given time. Using his terminology, psychologists have measured *interindividual* variation at a time rather than measuring *intraindividual* variation over time. This would not be a problem if interindividual variation was a reliable guide to intraindividual variation, but Molenaar argues that it is not. Except under stringent conditions, variation among the individuals in a population at a time will not necessarily reflect the variation within any of its members over time.

Molenaar argues that interindividual variation will only be a reliable guide for intraindividual variation if a condition known as *ergodicity* is met. Ergodicity obtains just in case as one collects more data on between-individual variation, the sample probability distribution asymptotically approaches that of the probability distribution for within-individual variation. In other words, in the limit, the probability distribution for the population and that for the individual are the same. Ergodicity fails either when the individuals in the population are heterogeneous (i.e., they have different probability distributions) or when the function modeling the variation is non-stationary over time. A function is non-stationary over time when the process generating the data changes with time. Molenaar (2004) suggests that developmental processes in psychology are non-stationary "almost by definition" (p. 215).

Ergodicity requires that the population and individual probability distributions be identical—rather than merely similar—and Molenaar does not prove that there is no weaker condition that would allow one to infer facts about an individual based on her population. For my purposes here, however, it is sufficient that Molenaar is clearly correct that variation in a population need not match the variation in any of its members. This is not an idiosyncratic feature of some sophisticated models, but is evident from simple examples. The fact that humans vary in their genetic code does not entail that there is any individual whose genetic code varies over time.

In the following, I argue that it is not possible for a factor to be a cause in a population, but not a cause in at least one of its members. This thesis may appear incompatible with Molenaar's thesis that between- and within-individual variation may come apart, for the following reason. In order to measure the causal effect of X on Y, there must be covariation between X and Y. If X and Y can covary among individuals in a population, without covarying within any of its members over time, doesn't it straightforwardly follow that X can cause Y in a population without causing Y in any of its members? In the next section I will show that it does not. The aim of the present section was to emphasize that I do not reject the claim that within-individual and between-individual *variation* can come apart. Rather, I argue that once one distinguishes between correlation and causation, it becomes evident that between- and within-individual causation cannot come apart in the same manner.

Table 1. Simpson's paradox case.

	E	~E		Recovery Rate
Combined				
Drug (C)	40	40	80	50%
No Drug (~C)	32	48	80	40%
	72	88	160	
Males				
Drug (C)	36	24	60	60%
No Drug (~C)	14	6	20	70%
	50	30	80	
Females				
Drug (C)	4	16	20	20%
No Drug (~C)	18	42	60	30%
	22	58	80	

Note. Adapted Pearl (2009). C = taking.

Lessons from Simpson's paradox

In this section, I analyze the well-known phenomenon of Simpson's paradox as a way of spelling out the distinction between causation and correlation. My discussion of the paradox will reveal that although there can be a correlation in a population that does not obtain in any of its subpopulations, there cannot be a cause in a population that is not a cause in any of its subpopulations. I derive this point within Pearl's (2009) causal modeling framework.[2] I then argue that within psychology, individuals can be characterized as subpopulations. It follows that there cannot be a cause in a population that is not a cause within any individual. From this claim, it is straightforward to show that there cannot be a between-subjects cause that is not a within-subjects cause.

Simpson's paradox refers to cases in which factor X raises the probability of factor Y in a population, but X does not raise Y's probability in any subgroup of that population. Pearl's (2009) example of Simpson's paradox involves a hypothetical study to determine whether a particular drug is effective. The population is divided into males and females. Within each group, participants who take the drug have a *lower* recovery rate than those in the control group. Within the combined population, however, taking the drug is positively correlated with recovery. A brief analysis of the data in Table 1 reveals what is going on. Regardless of whether they take the drug, males have a much higher recovery rate than females. Moreover, a greater number of males received the drug in the study. The information that one took the drug raises the probability that one is male and males are more likely to recover. It is for this reason that taking the drug is positively correlated with recovery in the population as a whole despite the fact that it is negatively correlated with recovery within any subpopulation.

As the numbers in Table 1 indicate, there is nothing paradoxical about a treatment being positively correlated with recovery in a whole population, but not in any subpopulation. The paradox only arises if one assumes that the positive correlation is due to a

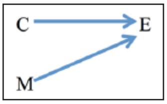

Figure 1. A causal graph indicating the causal relationships between variables *C*, *M* and *E*.

causal relationship. It *would* be paradoxical if the drug caused recovery in a population without being a cause of recovery in any subpopulation. A correlation between two variables can reflect either that one variable provides evidence for the presence of the other or that one is a cause of the other. In order to discover which causal relations obtain in a population, one needs to distinguish between the correlations that reflect causes from those that merely reflect evidential relevance. In Pearl's example, one needs to look at the male and female populations individually in order to correctly estimate the effect of the drug on recovery. In doing so, one "blocks" the evidential relationship between having taken the drug and having a property (i.e., being male) that predisposes one to recover and only the causal relationship remains.

The idea that correlation does not entail causation and that one must therefore distinguish between them is familiar. I appeal to Simpson's paradox to highlight how calling a relationship within a population "causal" places constraints on the possible causal relationships within that population's subpopulations. For example, I will show that in the model in Figure 1, the effect of *C* on *E* in the population is a weighted average of its effect on *E* in subpopulations stratified based on their value for *M*. The effect of *C* on *E* in the population constrains the possible effects within the subpopulations, since *C* cannot be a cause of *E* in the population without also being a cause in at least some of its subpopulations.

In Figure 1 there are no common causes or other confounders of the relationship between *C* and *E*. The correlation between *C* and *E* is therefore entirely due to the causal relationship between *C* and *E* and the effect of *C* on *E* is simply Pr(E|C). The effect of *C* on *E* within a population for which *M* has the value of *m* is Pr(E|C,M=m). The relationship between the effect in the population and that in the subpopulations is given by equation (A):

$$\text{(A)} \qquad \Pr\left(E\middle|C\right)=\sum_{\mathrm{m}}\Pr\left(E\middle|C,M\right)Pr\left(M\right)$$

This equation is a precise formulation of the claim that the effect of *C* on *E* in the population is a weighted average of the effect of *C* on *E* in subpopulations stratified using *M*.

To explicitly represent that the causal effects in Figure 1 are given by the conditional probabilities in (A), we can rewrite it as follows:

$$\text{(B)} \quad \Pr\left(E\middle|do\left(C\right)\right)=\sum_{\mathrm{m}}\Pr\left(E\middle|do\left(C\right),M\right)Pr\left(M\right)$$

The expression Pr(E|do(C)) refers to the probability of *E* given an *intervention* on *C*. An *intervention* on *C* sets *C* to a particular value so that it no longer depends on its causes other than the intervention.[3] An intervention need not reflect the activity of an agent; any exogenous source of variation in the value of *C* can be modeled as an intervention on *C*. In cases where the relation between *C* and *E* is not confounded, Pr(E|C)=Pr(E|do(C)), though this equality will not hold in general. Where the two probabilities diverge, the causal effect of *C* on *E* is given by Pr(E|do(C)), rather than Pr(E|C). As one cannot intervene on latent variables, talk of interventions may seem out of place here. Even if one cannot perform an intervention, however, it may still be possible to define a causal effect in terms of a hypothetical intervention. If one cannot even suggest a thought experiment in which one intervenes on *C* or if there are many possible interventions on *C* with different effects on *E*, then the effect of *C* on *E* is not well defined.[4]

Let's say that *C* causes *E in population P=p* just in case there are at least two values of *C*, c and c', such that Pr(E|do(C=c),P=p)≠Pr(E|do(C=c'),P=p). This inequality states that there is a possible intervention from c to c' that changes the probability of *E* in that population. Equation (B) reveals that the effect of *C* on *E* in the population as a whole must be an average of its effects in the subpopulations stratified based on *M*. This entails that it is impossible for *C* to cause *E* in the population as a whole, but not cause *E* in any of the subpopulations.[5]

In order to argue that it is impossible for there to be a cause in a population but not in any of its subpopulations, I need to consider all possible ways of dividing a population into subpopulations using a variable that is not on a path between *C* and *E*.[6] So far I have only discussed the case in which one stratifies the population based on *M* in Figure 1. *M* is an example of an *effect modifier* for the effect of *C* on *E*, since it is a cause of *E* that is not on the path between *C* and *E*. All effect modifiers are related to a cause of *E* that is not on the path from *C* to *E* (see Vanderweele & Robins, 2007, for a precise characterization). When one conditions on an effect modifier, the causal decomposition is the same as that given in (B) and, consequently, there cannot be a cause in the population that is not a cause in some of its subpopulations. When one stratifies a variable that is not an effect modifier (and not on a path between *C* and *E*), the effect of *C* on *E* in the resulting subpopulations is the same as that in the population.[7] Since such variables make no difference in the effect of *C* on *E*, there will be no difference in the effect of *C* on *E* in populations with different values for these variables.

To summarize, stratifying a population using an effect modifier will yield subpopulations in which *C* causes *E* in at least one subpopulation. Stratifying a population based on a variable that is not an effect modifier will yield subpopulations in which the effect of *C* on *E* is the same as that in the whole population. It is therefore not possible for there to be a cause in a population that is not a cause in any of its subpopulations.

The next step in the argument is from the claim that there cannot be a cause in a population, but not in any of its subpopulations, to the claim that there cannot be a cause in a population but in none of its members. Provided that individuals can be specified in terms of their values for a set of variables, any individual can be treated as a subpopulation (albeit possibly one with a single member). The premise that individuals can be treated as subpopulations presupposes that a psychological model for some individual, Ida, does not seek to explain why Ida answered a test question correctly on a particular day, but rather why an individual with all of Ida's properties tends to answer such questions correctly. I assume that this is the most one could expect from a psychological model, even a model for a particular individual. In saying that individuals can be characterized as subpopulations, I am not claiming that it is possible in practice to estimate the effect in those subpopulations. My point is a conceptual one about the relationship between a causal factor in a population and the influence of that factor within its members, not a practical point about our ability to measure the effects of that factor within individuals.

The claim that individuals can be characterized as subpopulations should not be confused with the claim that individuals consist of an aggregate of entities having the same properties as that individual. Individuals are representable as populations insofar as they are characterized in terms of their properties. If Ida, and only Ida, has properties P_1, P_2, …, P_n, the causal model for Ida will specify how someone with those properties will respond to various treatments. The model characterizes a population in the sense that it represents *any* hypothetical individual with Ida's properties.

So far I have argued for two claims. First, it cannot be the case that *A* causes *B* in the population as a whole but not in any subpopulation. Second, individuals in psychology can be characterized as subpopulations. It follows that *A* cannot cause *B* in a population, but not in any of its members. Given that there cannot be a cause in a population but in none of its members, it is straightforward to show that there cannot be a causal variable with a between-subjects interpretation, but no within-subjects interpretation. If a variable *X* has an exclusively between-subjects interpretation, then *X* will be a cause in a population in which individuals differ in a value of *X* without its being a cause in any member of that population. As this is impossible, *X* (or any variable) cannot be exclusively a between-subjects cause.

The argument for the between-subjects interpretation

If there cannot be an exclusively between-subjects cause, then Borsboom et al.'s claim that intelligence (as typically measured by psychologists) is such a cause must be mistaken. In this section, I examine their argument that intelligence is a cause with an unproblematic between-subject interpretation. In the following section, I will consider their argument for why intelligence cannot be a within-subject cause. I will largely ignore the issue of whether latent variable models should be interpreted realistically. My thesis is a conditional one: if intelligence is a cause, it is a within-subjects cause.

The authors defend the between-subjects interpretation of intelligence on the grounds that it meets Mill's (1843) three criteria for causation:

> X can be considered a cause of Y if (a) X and Y covary; (b) X precedes Y; and (c) ceteris paribus, Y does not occur if X does not occur. (Borsboom et al., 2003, p. 211)

Mill's account has some well-known problems. Suppose that *X* and *Y* share a common cause *C* and that *X* comes before *Y*. Conditions (a) and (b) clearly apply. Here's an argument that (c) does as well. Had *X* not happened, we would be (or, if there were other causes of *X*, would be more likely to be) in a situation in which *C* did not happen either, but if C did not occur, then neither would Y, ceteris paribus. So all three conditions obtain, even though X is not a cause of Y.[8] The authors are not, however, committed to Mill's view and claim (without providing details) that it is possible to give a between-subjects interpretation of the latent variable using Lewis' (1973) or Pearl's (2009) account.

It is worth thinking a bit more about why Mill's account is inadequate. In the case where X and Y share a common cause, the presence of Y provides evidence for X even though X does not cause Y. Condition (c), however, is unable to distinguish between cases where X is a cause of Y and cases where X merely provides evidence for Y. It therefore cannot provide a method for distinguishing between causal and non-causal relations. Thus, Borsboom et al.'s argument that the latent variable for intelligence meets all three conditions does not establish that it is a cause, whether of the between-subjects or within-subjects variety.

Pearl's account does not require condition (b) and it replaces condition (c) with the condition that the probabilistic relationship between X and Y is invariant to interventions on X. Since I rely on Pearl's framework for my proof in the previous section, one avenue for avoiding my conclusion is to reject his framework. Aside from the fact that Borsboom et al. claim that their argument *is* compatible with it, the substantive question is whether there is a framework in which there can be an exclusively between-subject cause. I take it as given that causation consists of more than just correlation plus time ordering. It remains to be seen that there is an additional condition other than Pearl's that provides both an adequate account of causation and allows for exclusively between-subjects causes.

The argument against the within-subjects interpretation

Borsboom et al.'s main argument for the claim that a latent variable cannot be a within-subjects cause is that since each subject has a single unvarying level of intelligence (at a time), the variable will have a constant value for each subject and a constant cannot be a cause. As the authors put the point: "If there is no covariation, there cannot be causation" (2003, p. 211). This preliminary assertion rests on dubious grounds. The epistemological point that if a variable does not vary in our data set we cannot *determine* whether it causes another variable does not settle whether it *does* cause another variable. Moreover, as they go on to discuss, this problem is avoidable if one interprets causal claims as counterfactuals. On this interpretation, whether intelligence is a within-subjects cause is not settled by the actual value of an individual's intelligence variable, but by a counterfactual concerning whether that individual would have done better on an exam had she had a higher level of intelligence.

So what is wrong with the counterfactual "had Einstein been less intelligent, he would have done less well on the IQ test"? The authors reformulate this as saying that had Einstein had the intelligence level of John, he would not have done as well and claim that although this counterfactual appears to treat intelligence as a within-subjects cause, appearances are deceiving. This counterfactual is not about whether Einstein would have been able to score as highly had he been less intelligent, but rather, whether he would have been able to do so had he had John's level of intelligence. Regarding this latter counterfactual they ask "But does this counterfactual statement express the causal efficacy of intelligence within Einstein?" (Borsboom et al., 2003, p. 212) and suggest that it only expresses the between-subjects counterfactual that Einstein did better than John because he was more intelligent.

To dramatize their point, Borsboom et al. consider a case in which there is a counterfactual about Einstein that they claim *must* be interpreted as a between-subjects counterfactual: "had Einstein been a fruit fly, then he would have been less intelligent." They write:

> However, the counterfactual statement is completely unintelligible except when interpreted as expressing knowledge concerning the difference between human beings (a population) and fruit flies (another population). It does not contain information on the structure of Einstein's intellect and much less on the alleged causal power of Einstein's intelligence. It contains only the information that Einstein will score higher on an IQ test than a fruit fly because he is more intelligent than a fruit fly—but this is exactly the between-subjects formulation of the causal account. (2003, p. 212)

If we interpret the fruit fly counterfactual in the way that the authors say we should, the intelligence variable that is the subject of the counterfactual is not a cause. There is no problem with positing a variable that is correlated with the performance of members of each population on an IQ test. For this to be a *causal* variable, being in a population with a certain value for that variable needs to not merely provide evidence for one's performance, but it must be the case that one's being in one population or another makes a difference in whether one does better or worse on the IQ test, on average. A variable that does nothing other than express "knowledge concerning the difference" between two populations is not a causal variable.

Perhaps the fruit fly example is problematic for reasons that do not apply to more commonplace counterfactuals about intelligence, such as the one comparing Einstein to John. Part of the problem with comparing Einstein to a fruit fly is that we have no idea how one could change his intelligence to that of a fruit fly without changing him into an individual that is no longer Einstein. The idea that this is problematic motivates Holland's (1986) claim that essential attributes of individuals are not causes.[9] It also underlies Woodward's (2003) assertion that all causal explanations concern "same-object" counterfactuals. Both Holland and Woodward would say that a counterfactual with the antecedent "had Einstein been John" is not a causal counterfactual. For a counterfactual about Einstein's intelligence to be causal, there must be a conceivable "surgical" intervention by which one changes his intelligence while keeping his other properties fixed. If you replace Einstein with John, you replace all of his properties, so such a replacement is not a surgical intervention. In contrast, a causal counterfactual about Einstein's intelligence might describe the effects of a surgical intervention on Einstein that changes his intelligence so that it equals John's. If such an intervention is possible, then it clearly is the case that intelligence has a within-subjects interpretation. The intervention, being surgical, describes the effect of changing *Einstein's* intelligence so that it equals John's. It is not a counterfactual about the intelligence level that Einstein would have had had he *been* John. The effect in Einstein of having John's level of intelligence need not equal the effect in John of having John's level of intelligence.

This discussion of interventions allows us to explicate Borsboom et al.'s notion of "exchangeability." The authors note that with height, unlike intelligence, one can infer that any between-subjects differences in the effects of height will correspond to within-subjects effect of changing the height of an individual. The difference between height and intelligence is that different individuals are *exchangeable* with regards to height, but not with regards to intelligence. Borsboom et al. explain: "If we give Einstein platform shoes that make him taller, he is, in all relevant respects, exchangeable with the taller person in the example" (2003, p. 212). This example intuitively captures the idea that it is possible to intervene on Einstein's height in such a way that his other properties are not affected by the intervention. If exchangeability is cashed out in terms of the possibility of performing an intervention, one cannot use exchangeability to distinguish between different types of causes. If there were no conceivable intervention on Einstein's height, height would not be a cause and thus would not be a between-subjects cause. Exchangeability does not allow one to distinguish between different types of causes; the test for whether a variable is a cause is whether individuals are exchangeable with respect to it.

Part of the alleged problem with comparing Einstein to John is that the variable for intelligence might indicate entirely different processes in each individual. According to Borsboom et al.:

> the between-subjects latent variable will not indicate the same process in each subject. Therefore, the causal agent (i.e., the position on the latent variable) that is posited within subjects on the basis of a between-subjects model does not refer to the same process in all subjects. (2003, p. 214)

They give an example in which one person fails to answer an IQ test question because of a deficiency in memory and another misses the question due to a deficiency in spatial reasoning. To evaluate whether the existence of heterogeneous processes responsible for test performance threatens the within-subjects interpretation of intelligence, we need to think more about the relationship between these processes and the intelligence variable. A natural suggestion is that intelligence is a causal variable that is *realized* differently within different individuals. The authors suggest that the processes within individuals realize the between-subjects latent variable (Borsboom et al., 2003, p. 216). Yet, if intelligence is realized in different individuals, then it clearly is a within-subjects cause. This would be analogous to a case in which causing pain either to an octopus or a human will lead them to express aversive behavior, even though the pain mechanisms and the aversive behaviors might be very different. The fact that pain is realized differently in humans and octopuses, presents no threat to the claim that pain is a cause of aversive behavior within members of each species.

In spelling out why the heterogeneity of individual processes is a problem for the within-subjects interpretation, Borsboom et al. claim that the processes responsible for IQ scores are "qualitatively" different (2003, p. 214). By this they mean that the between-subjects variable for intelligence picks out a set of processes that have nothing in common other than that they are causes of IQ test performance. If, however, the process by which one individual answers IQ test questions is totally different from that by which another one does, what could it mean to say that there is a single causal variable called intelligence that explains why these two individuals performed differently on the test? Saying that there is a causal variable with only a between-subjects interpretation is unhelpful. While there can be a non-causal variable that is correlated with IQ test performance and which is thus useful for predicting how an individual will do on a test, more is required in order to show that there is a *cause* of the variation in performance. The claim that the individual processes corresponding to intelligence differ radically across individuals does not shed any light on how there might be a cause with an exclusively between-subjects interpretation.

In fact, if there are many different processes corresponding to intelligence, this suggests that there are numerous distinct ways to intervene on intelligence and which intervention one chooses makes a difference in the effects of intelligence. If so, then intelligence is not a well-defined causal variable.[10] Unless one specifies which method one uses to intervene on "intelligence," questions about the effects of intelligence are ill posed. Of course, even if intelligence corresponds to a well-defined intervention, it may have different effects in different individuals. Einstein and John might differ in properties that are modifiers for the effect of intelligence on test performance. As we have seen in an earlier section, the effect in the population of Einstein and John will then be the *average* of the effect in each of them. While the effect in the population is an average effect, it is the average effect of performing the *same* intervention in all members of the population.

I have suggested that two subjects are exchangeable with regards to intelligence just in case there is a possible intervention on the intelligence of one of the subjects so that she has the same intelligence level as the other. Although I have argued that the concept of exchangeability cannot ground the distinction between within-subjects and between subjects-causes, it is worthwhile to consider Borsboom et al.'s explanation for why individuals are not exchangeable with regards to intelligence, as measured by standard psychological models:

> such an argument [for the claim individuals are exchangeable with regards to intelligence] would have to rest on the idea that, by necessity, there has to be a strong relation between models for within-subjects variability and models for between-subjects variability. It turns out that this idea is untenable because there is a surprising lack of relation between within-subjects models and between-subjects models. To discuss within-subjects models, we now need to extend our discussion to the time domain. This is necessary because to model within-subjects variability, there has to be variability, and variability requires replications of some kind; moreover, if variability cannot result from sampling across subjects, it has to come from sampling within subjects. (Borsboom et al., 2003, p. 212)

The argument here relies on the authors' earlier claim that there can be no causation without covariation. Since there can be no causation without covariation and intelligence models treat individual intelligence as fixed at time, any within-individual causal claims must concern the way that an individual varies over time. The Molenaar (2004) results show, however, that variation in a population will not be a reliable guide to the way that an individual varies over time.

Contrary to what the authors say, within-subjects variation is not required for there to be within-subjects causation. We can see this by considering the methodology of randomized control trials. In the cases that the authors are discussing the causal variables are latent, so no such trial is possible (unless we measure them), but what I will say about the causal effects discovered in randomized trials extends to latent causes as well. Within a randomized trial, every subject either receives the treatment or is in the control group, so there is no within-subjects variability with regards to treatment.[11] This does not mean, however, that randomized trials do not measure within-subjects effects. In the limit, randomization ensures that whether an individual receives the treatment does not depend on any factors that might influence the causal effect of the treatment on the outcome. As a result, the average difference between the treatment and control group provides an estimate of the average causal effect of giving the treatment to all of the members of the population. Even though one never observes how the members in the treatment group would have been affected had they not been treated or how the members of the control group would have been affected had they been treated, one can still determine how much the treatment influences an individual's outcome, on average.

Pearl's analysis of Simpson's paradox provides an explanation for why randomization works. Prior to randomization, any correlation between treatment and outcome might reflect either that the treatment affects the outcome or that knowing that someone received the treatment provides evidence that she would be more likely to recover regardless of whether or not she were to receive the treatment. Successful randomization ensures that someone's receiving the treatment provides no evidence regarding how she might react to it. The treatment only depends on the randomizing device, which by design is not correlated with any factor responsible for a subject's outcome. Consequently, any correlation between the treatment and the outcome must be causal.

It might appear that because randomized control trials only estimate an average effect, it is incorrect to say that they estimate a within-subjects effect. Yet, an average within-subjects effect is still a within-subjects effect. Were one to gain more information about an individual's effect modifiers, one would be able to make a more reliable prediction about how that individual would respond to a treatment. But the fact that the average effect is not maximally precise does not detract from its being a within-individual effect. A randomized trial estimates how an individual of a certain type would respond to treatment on average, not merely how an individual differs from individuals of a different type.

To be clear, I am not denying that there is an important question regarding what we can infer about how an individual would react to a treatment based on a randomized control test on a population. Rather, I am pointing out that this is not the question posed by Borsboom et al. (2003). For them, the problem is that there is no within-subjects effect in a population in which there is no within-subjects variation. Our discussion of randomized control trials reveals this claim to be incorrect. Even though there are many cases in which randomized control trials are unfeasible, the fact that such trials measure a within-subjects causal effect in populations with no within-subjects variation in treatment is a decisive counterexample to the authors' claim that within-subjects causation requires within-subjects variation.

The idea that average causal effects have within-subjects meaning reflects my earlier point that while a population-level positive correlation between variables need not reflect anything about the correlations within subpopulations, a population-level *cause* does place constraints on what causal relations might obtain within subpopulations. Borsboom et al.'s claim that "the standard measurement model [for intelligence] has virtually nothing to say about characteristics of individuals" (2003, p. 206) is incompatible with their thesis that this model should be interpreted as a causal model.

Towards the end of their argument against the within-subjects interpretation of intelligence, Borsboom et al. remark:

> The main reason for the failure of the within-subjects causal account seems to be that it rests on the misinterpretation of a measurement model as a process model, that is, as a mechanism that operates at the level of the individual. (2003, p. 214)

They then proceed to give other examples of similar mistakes, including the following:

> [A] linear model for interindividual variations in height does not imply that individual growth curves are linear; that 30% of the interindividual variation in success in college may be predicted from the grade point average in high school does not mean that 30% of the exams you passed were predictable from your high school grades. (p. 214)

As well as this fallacy:

> that there is a sex difference in verbal ability does not mean that your verbal ability will increase if you undergo a sex change operation. (p. 214)

The authors consider these to be all instances of the same fallacy, but they are actually distinct. The fallacy in the first examples is that of inferring that the variation among variables in a population corresponds to the variation of those variables within an individual over time. This is the fallacy that Molenaar (2004) warns against. The fallacy in the second passage, however, is one of inferring causation from correlation. I find it hard to imagine that many psychologists commit this particular fallacy, since even proponents of the view that sex is a cause of differences in verbal ability presumably do not think that the causal mechanism runs through the genitalia. This detail aside, the intended fallacy here is one of moving from a claim about the correlation between sex and ability to a causal claim about the results of intervening on sex.

The authors' failure to distinguish between causal and statistical fallacies underlies their problematic claim that intelligence can be exclusively a between-subjects cause. If intelligence is a cause, it is a within-subjects cause. The appearance that Molenaar's (2004) results show otherwise is an illusion resulting from not properly distinguishing between causal and non-causal models.

Discussion

Borsboom et al. (2003) are correct that the magnitude of a causal effect in a population is an unreliable guide to that cause's effect within any particular individual. We should therefore be cautious in applying models for the effects of intelligence in a population to its members. It is important to understand *why* population-level models might not apply to particular individuals. According to the authors, there can be a cause within an individual only if there is variation within that individual. Standard intelligence models cannot be applied to individuals in principle, since they treat individuals as having a fixed level of intelligence at a time. In contrast, I have argued that what determines whether there is a cause within an individual is not whether the individual *actually* varies, but rather whether it would be possible to induce variation within the individual via a hypothetical intervention. The reason one should not assume that a cause in a population has a similar effect in an individual is not that the causal effect in the population has no within-individual meaning. Rather, it is because the effect in the population is only an average effect and individuals may deviate from the average.

On Borsboom et al.'s (2003) approach, learning that intelligence has an effect in a population provides no reason for thinking that it will have a similar effect within an individual. I reply that since an effect in a population is an average of that effect in its members, learning that intelligence has an effect in a population provides *some* evidence for believing that it has a similar effect in a particular member. On the assumption that the effects of intelligence vary greatly among individuals, establishing that intelligence has a certain effect in the population would provide only weak support for the claim that it has a similar effect in an individual. If, instead, we assume that the effect of intelligence does not vary greatly among individuals, then the effect of intelligence in a population will be a better guide to its effect within individuals. To address the problem of extrapolating from a population to an individual, we should not discard the information learned from populations. What is required is an account of when the effect in a population provides stronger or weaker evidence for the effect within a given individual.

Acknowledgements

I am very grateful to the following people for their valuable feedback and advice: Dan Hausman, Denny Borsboom, Peter Steiner, Olav Vassend, Hayley Clatterbuck, Shannon Nolen, Reuben Stern, and an anonymous reviewer.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Notes

1. See Danzinger (1990) and Lamiell (2003) for similar arguments.
2. Although I rely on Pearl here, his modeling framework is very similar to that of Spirtes, Glymour, and Scheines (2000).
3. Several authors (Eberhardt & Scheines, 2007; Korb, Hope, Nicholson, & Axnick, 2004) have suggested expanding the definition of "intervention" to include changes that do not *determine* the value of the variable on the receiving end of the intervention, but merely alter its probability distribution. Such "soft" interventions would not fully disrupt the influence between a variable and its causes other than the intervention. My preference is to define "intervention" so that interventions do determine the value of a variable and to model "soft interventions" on X as interventions on an exogenous cause of X.
4. See Note 8 for further discussion.

5. Proof: Suppose, for reductio, that *C* causes *E* in the population, but not in any subpopulation stratified using *M*. From the claim that *C* does not cause *E* in any subpopulation and the definition of *cause in a population*, it follows that in every population *M=m* and for any two values of *C, c* and *c'*:

$$\Pr\left(E \mid do(c), M = m\right)\Pr\left(M = m\right) = \Pr\left(do(c'), M = m\right)\Pr\left(M = m\right)$$

Accordingly,

$$\sum_m \Pr\left(E \mid do(C = c), M\right)\Pr\left(M\right) = \sum_m \Pr\left(E \mid do(C = c'), M\right)\Pr\left(M\right)$$

From (B), it follows that for all values of C:

$$\Pr\left(E \mid do(C = c)\right) = \Pr\left(E \mid do(C = c')\right)$$

Which contradicts the supposition that *C* causes *E* in the population as a whole.

6. Variables on a path between *C* and *E* are called *mediators.* If there is a mediator on the sole path between *C* and *E* and one stratifies a population on it, then *C* will not have any effect on *E* within the resulting subpopulations. In determining the effect of *C* on *E*, one should not condition on any mediators (Pearl, 2009, p. 178) so I exclude them from the present discussion.
7. Of course, stratifying based on a non-modifier may yield a subpopulation where the *correlations* between *C* and *E* do not identify the effect of *C* and *E* in the population (even if there is no confounding). For example, if you stratify based on a common effect of *C* and *E*, this produces endogenous selection bias (Elwert & Winship, 2014). Accordingly the correlation between *C* and *E* will vary among the strata. It is important to see that this variation does not reflect a difference in the *effect* of *C* on *E* among the strata, but only that the correlations between *C* and *E* in the strata do not identify any causal effect.
8. One might argue that the counterfactual I gave to argue that (c) obtains is somehow not a "proper" counterfactual (in Lewis', 1973 terminology, it "backtracks"), but to make this argument one would need to specify which factors are held equal when we evaluate a ceteris paribus clause, and the authors do not do so.
9. The authors incorrectly attribute to Holland (1986) the view that "a constant cannot be a cause" (Borsboom et al., 2003, p. 211). Holland's claim, however, is not that no constant can be a cause, but rather that attributes such as sex and race cannot be causes, since one cannot intervene upon these properties in an individual in such a way that the individual intervened upon remains the *same* individual. The alleged problem with treating sex and race as causes is not that they happen to have constant values, but that it is impossible to intervene upon them. See Marcellesi (2013) for a response to Holland.
10. Within Pearl's framework, it is a theorem that if a causal graph is an adequate representation of a set of variables, interventions on a variable will only influence that variable's effects through influencing that variable (Pearl, 2010). This entails that if there are multiple possible interventions changing causal variable *C*'s value from *c* to *c', C*'s effect on *E* will be the same regardless of which intervention one chooses. It follows that if different interventions on *C* make a difference in *C*'s effect on *E, C* is not a well-defined causal variable. In Rubin's potential outcomes framework (Rubin, 1974), the invariance of causal effects to different interventions on the cause is not a theorem, but an assumption. This assumption is called *consistency.* Although there have been recent debates about how to understand consistency (Cole

& Frangakis, 2009; Hernán & VanderWeele, 2011; Pearl, 2010; VanderWeele, 2009), all parties to the dispute agree that causal variables must correspond to well-defined interventions.

11. One must not confuse the entirely uncontroversial claim that there is no within-individual variation in the treatment with the presumably false claim that there is no variation in how individuals respond to the treatment.

References

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review, 110*(2), 203–219.

Cole, S. R., & Frangakis, C. E. (2009). The consistency statement in causal inference: A definition or an assumption? *Epidemiology, 20*, 3–5.

Danzinger, K. (1990). *Constructing the subject*. Cambridge, UK: Cambridge University Press.

Eberhardt, F., & Scheines, R. (2007). Interventions and causal inference. *Philosophy of Science, 74*(5), 981–995.

Elwert, F., & Winship, C. (2014). Endogenous selection bias: The problem of conditioning on a collider variable. *Annual Review of Sociology 2014, 40*, 31–53.

Hernán, M. A., & VanderWeele, T. J. (2011). Compound treatments and transportability of causal inference. *Epidemiology, 22*, 368–377.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*, 945–959.

Korb, K. B., Hope, L. R., Nicholson, A. E., & Axnick, K. (2004). Varieties of causal intervention. In C. Zhang, H. W. Guesgen, & W. K. Yeap (Eds.), *PRICAI 2004: Trends in artificial intelligence* (pp. 322–331). Berlin, Germany: Springer.

Lamiell, J. T. (2003). *Beyond individual and group differences: Human individuality, scientific psychology, and William Stern's critical personalism*. Thousand Oaks, CA: Sage.

Lewis, D. (1973). *Counterfactuals*. Oxford, UK: Blackwell.

Marcellesi, A. (2013). Is race a cause? *Philosophy of Science, 80*(5), 650–659.

Mill, J. S. (1843). *A system of logic*. London, UK: Oxford University Press.

Molenaar, P. C. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement, 2*(4), 201–218.

Pearl, J. (2009). *Causality: Models, reasoning and inference* (2nd ed.). Cambridge, MA: MIT press.

Pearl, J. (2010). On the consistency rule in causal inference: Axiom, definition, assumption, or theorem? *Epidemiology, 21*(6), 872–875.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66*, 688–701.

Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, prediction, and search* (Vol. 81). Cambridge, MA: The MIT Press.

VanderWeele, T. J. (2009). Concerning the consistency assumption in causal inference. *Epidemiology, 20*(1), 880–883.

VanderWeele, T. J., & Robins. J. M. (2007). Four types of effect modification: A classification based on directed acyclic graphs. *Epidemiology, 18*(5), 561–568.

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

Author biography

Naftali Weinberger is a graduate student in philosophy at the University of Wisconsin, Madison. He is currently researching the question of when one can infer that a causal relationship found to obtain in one population obtains in others as well.