November 2004

Should Causality Be Defined in Terms of Experiments?

David A. Kenny Charles M. Judd

University of Connecticut University of Colorado

The research was supported in part by grant from the National Institute of Mental Health (RO1-MH46567) to the first author.

Abstract

The paper reviews what has come to be called the Rubin-Holland model of causation. That model takes as its starting point experimentation and defines causation as a hypothetical value: a difference score between the observation of a unit in control and experimental condition. We discuss several different implementations of the model that Rubin and Holland have suggested. We present several other possible implementations with special emphasis of quasi-experiments. Finally, we note that causality can alternatively be better defined in terms of the absence of spuriousness.

Should Causality Be Defined in Terms of Experiments?

A central but vexing quest in psychological research is the goal of gathering data that permit causal inferences about the effect of one construct on another. Much has been written in the literature on research methods about procedures that facilitate such inferences (e.g., Campbell & Stanley, 1963; Cook & Campbell, 1979; Judd & Kenny, 1981; Reichardt, 2003; Shadish, Cook, & Campbell, 2002; West, Biesanz, & Pitts, 2000). The goal of this literature has been to identify the conditions for establishing causal relations.

In this quest, much attention has recently focused on what has come to be known as the Rubin–Holland Causal Model (RHCM; Holland, 1986; 1988; Rubin, 1980; 1986; 1990; 1991). Donald Rubin and Paul Holland have importantly established a basic vocabulary for defining and estimating causal effects and their model is being increasingly accepted in statistics, medicine, and psychology (e.g., West et al., 2000; Reichardt, 2003). This acceptance is in part due the elegance of the model and its links to what is clearly the strongest design for causal inference, the randomized experiment. Nevertheless, it is our contention that RHCM, for all its virtues, has problematic aspects and we would do well to expand our thinking about the variety of conditions that permit causal inference. As shown in the paper, we are not primarily taking issue with many of the specifics of RHCM. We agree that causal inference is possible using the model and that the implementations of the model that they have elaborated are appropriate. Nevertheless, we believe that there are other possibilities that are not contained in the model and that need to be articulated.

We begin by presenting the basics of RHCM. We then elaborate on how its authors have considered implementing the model, thus exploring the conditions under which they believe that causal inference is possible. Having done this, we then turn to a number of other conditions, not considered by Rubin and Holland, in which we believe causal inferences are possible. Our primary goal here is not to disagree with the basics of the model, but rather to extend our thinking about possible ways to estimate causal effects. In the final part of the paper, we consider alternatives to the model.

The Model

The clearest exposition of the model, we believe, is that set forth by Paul Holland in his 1986 article, "Statistics and Causal Inference" which appeared in the *Journal of the American Statistical Association*. He starts by defining $\boldsymbol{U}$ as a population of units which are the basic objects of study and which are expected to manifest the causal effect of some variable. Normally, in much of psychological research these units are human participants. Any particular unit in this population is designated as $\boldsymbol{u}$. He then defines three variables on which $\boldsymbol{u}$ is measured: $\boldsymbol{Y}$, $\boldsymbol{A}$, and $\boldsymbol{S}$. Both $\boldsymbol{Y(u)}$ and $\boldsymbol{A(u)}$ are measurements taken on unit $\boldsymbol{u}$, the former designated as a response variable, expected to manifest the causal impact of the independent variable, if there is one, and the latter an attribute of the unit, representing a property or characteristic of the unit that is assumed to be unchanging. The term $\boldsymbol{S(u)}$ represents the independent variable whose causal effect is of interest. For the sake of simplicity, it is assumed to have only two possible values, designated as either $\boldsymbol{S} = \boldsymbol{t}$ or $\boldsymbol{S} = \boldsymbol{c}$. The choice of these values is perhaps unsurprising, with $\boldsymbol{t}$ representing some treatment or intervention and $\boldsymbol{c}$ representing a control condition. What is important is that any given unit could theoretically be observed in either of these levels of $\boldsymbol{S}$. The variable $\boldsymbol{A}$ is, however, an unchanging characteristic of $\boldsymbol{u}$. According to Holland, causal inferences can only be made in the case of $\boldsymbol{S}$ variables, not for $\boldsymbol{A}$ variables.

Holland then defines $\boldsymbol{Y_t(u)}$ as the value of the response that would be observed if the unit were exposed to $\boldsymbol{t}$ and $\boldsymbol{Y_c(u)}$ as the value of the response that would be observed if the unit were exposed to $\boldsymbol{c}$. It is the difference between these two theoretical values that is the causal impact of the independent variable, according to the model:

$$\boldsymbol{Y_t(u) - Y_c(u)}$$

He defines this as the causal impact of treatment $\boldsymbol{t}$ on unit $\boldsymbol{u}$.

There are two implicit assumptions of the model at this point. The first is that any possible causal effect is possible only when the *same* unit is in theory observable in each of the different levels of the variable whose causal impact is to be assessed. Because it is logically impossible for the same unit to simultaneously be in two different conditions, the measure of causal effect is a counterfactual. The second assumption is that that causal impact is assessable at the level of the individual unit.

Although not emphasized by Holland, Rubin (1986) had described the stable-unit-treatment-value assumption or SUTVA. Rubin (1986) define SUTVA as "the a priori assumption that the value of $\boldsymbol{Y}$ for unit $\boldsymbol{u}$ when exposed to treat $\boldsymbol{t}$ will be the same no matter what mechanism is used to assign treatment $\boldsymbol{t}$ to unit $\boldsymbol{u}$ and no matter what treatments the other units receive" (p. 961).

Given this model, Holland outlined what he called "the fundamental problem of causal inference: It is impossible to *observe* the value of $\boldsymbol{Y_t(u)}$ and $\boldsymbol{Y_c(u)}$ on the same unit, and therefore, it is impossible to *observe* the effect of $\boldsymbol{t}$ on $\boldsymbol{u}$" (p. 947; italics in original). Given this, he suggested that causal inference can be done only if one makes certain sets of untestable assumptions that underlie the alternative ways in which the RHCM might be implemented. The goal of these implementations is to develop an estimate of the causal impact of ***t*** on ***u*** that would equal the theoretically defined value if the assumptions underlying the implementation were met.

In the following sections, we outline the various implementation possibilities that Rubin and Holland have considered, and the assumptions that they argued underlie each. We then suggest that there are additional implementations that they fail to consider and that have the potential for considerably extending the variety of situations in which causal inference is possible. However, before doing this, we raise several theoretical issues about the model, in its theoretical rather than its implemented form.

From our perspective, the distinction between ***S*** and ***A*** variables is an unclear and, we feel, unnecessary one. Holland seems to believe that some variables are potentially manipulable and others not, in the sense that for ***S*** variables the same unit could be observed in theory under both its values, while for ***A*** variables this could not be the case. And, according to the RHCM, it is possible to speak of causal effects only in the case of ***S*** variables.

The classic ***A*** variable might be a unit's gender. Up until very recently, it was not possible to vary the gender of a unit; it was a stable attribute of the unit and the measurement of ***Y*** under its alternative levels could not be contemplated. According to the RHCM, then, it would seem that the question of the causal impact of gender on some response variable could not possibly be addressed. More recently, gender has become manipulable. So what used to be an ***A*** variable, whose causal impact could never be assessed, now becomes a potential ***S*** variable. It seems peculiar that the plausibility of gender having a causal impact on some response variable that might potentially be estimable should vary as a function of the development of surgical and hormonal treatments.[1] Thus, we do not see the utility of the ***A/S*** distinction.

A second theoretical issue concerns the assumption, which we have already pointed to, that there is a single causal effect, in theory manifested at the level of the individual unit. If this were the case, then the magnitude of that causal impact would seem in theory to be the same for all units, manifested by them all to the same degree. It seems to us that this is a surprising and unnecessary assumption. While it is part of RHCM, there seems little reason to believe that it is a necessary assumption in attempting to make causal inferences. The causal impact of an independent variable may certainly vary across units and across occasions and so the causal effect could be defined as the average effect in the population of units ***U***.

A third implicit assumption of RHCM that follows from the above is that units are completely interchangeable. In some sense they are disembodied; anyone is as good as any other one for purposes of assessing the true causal effect of some variable. But this seems surprising indeed given all we know about the complexities of human behavior and the role of contextual factors in influencing behavior. It seems obvious to us that units are not interchangeable; rather they are embedded in contexts and theories that fail to realize this seem to us incomplete.

Finally, another related assumption of RHCM is that the ***Y*** is measurable without error. Different units, even if they could be observed in both levels of ***S***, might manifest very different causal effects, both because those causal effects vary systematically across units and also because the measurement of $\boldsymbol{Y}$ is imperfect, varying across occasions and units randomly.

Implementations of the Rubin–Holland Causal Model

The literature devoted to RCHM has emphasized five different ways by which the effect of $\boldsymbol{t}$ on $\boldsymbol{u}$ might be observed or estimated, under certain assumptions. These procedures can overcome what Holland (1986) refers to as the "fundamental problem of causal inference" through the specific assumptions made in each implementation. The methods are repeated measures, unit homogeneity, randomization of units, observational studies, and treatment compliance studies.

Repeated Measures

The first implementation that Holland (1986) discusses involves repeated observations of unit $\boldsymbol{u}$ in the different levels of $\boldsymbol{S}$. Suppose that for unit $\boldsymbol{u}$, we first measure $\boldsymbol{Y_c(u)}$ and then we subsequently measure $\boldsymbol{Y_t(u)}$. Under two assumptions, outlined by Holland (1986), the resulting difference tells us about the causal effect of $\boldsymbol{t}$ on $\boldsymbol{u}$. The first assumption is *temporal stability*: "the value of $\boldsymbol{Y_c(u)}$ does not depend on *when* the sequence 'apply $\boldsymbol{c}$ to $\boldsymbol{u}$ then measure $\boldsymbol{Y}$ on $\boldsymbol{u}$' occurs" (p. 948, italics in the original). And the same assumption holds for treatment $\boldsymbol{t}$. The second assumption is called *causal transience*: the value of $\boldsymbol{Y_t(u)}$ is not affected by the prior exposure of the unit to $\boldsymbol{c}$.

We certainly agree with Holland that if these two assumptions can be met, then causal inference from the repeated measures implementation is possible. However, we would believer that these two assumptions are very unlikely ever to be met, especially within psychology. Particularly the first one seems totally implausible, at least when the units are humans or some other animate entity. Units can be expected to change over time; they mature, there are historical events that intervene that affect $\boldsymbol{Y}$, and there are spontaneous, stochastic changes that occur. Even if we were to grant that the assumption of causal transience can be met for some causes, satisfying the temporal stability assumption seems almost always to be implausible. There must be other solutions to causal inference in repeated measures contexts, and we point to an obvious one below, one that appears never to have been considered within the confines of RHCM.

Unit Homogeneity

The second implementation that Holland considers is that of *unit homogeneity*. If one makes the assumption that " $\boldsymbol{Y_t(u_1) = Y_t(u_2)}$ and $\boldsymbol{Y_c(u_1) = Y_c(u_2)}$ for units $\boldsymbol{u_1}$ and $\boldsymbol{u_2}$" then the causal effect equals $\boldsymbol{Y_t(u_1) - Y_c(u_2)}$. In other words, a comparison between two units can be used to estimate the causal impact on any one unit. It seems to us that this unit homogeneity assumption is implicit in RHCM itself, because, as we have said, the model assumes that the treatment effect is identically defined for each individual unit.

Like the temporal stability assumption in the repeated measures implementation, however, it seems to us very unlikely that the unit homogeneity assumption is ever met with animate units. Because of a host of genetic and contextual variables, no two animate units are the same. Additionally, as we have already suggested, treatment effects are likely to vary across units, both systematically and because of random errors of measurement. In sum, the implementation of the model that assumes unit homogeneity is implausible in psychological research.

Randomization of Units

Because *$Y_t(u)$* and *$Y_c(u)$* cannot be observed for each person, we might randomly assign each unit to one of the two conditions and then observe a set of units in each of the two conditions. The difference between the treatment condition mean and the control group mean provides an estimate of the causal effect of the treatment. Because each unit has the same probability of being in a given condition, the expectation of the difference between these two means is causal effect of the treatment.

There are several assumptions made for the randomization of units. The first is that the assignment process is known to be random. Haphazard assignment is not likely to be random and so should not be presumed to be random. Holland (1986) uses the term "physical randomization" (p. 948) to convey the idea that a known random process is used (e.g., dice rolled or numbers taken from a table of random numbers).

The second is that the units be independent of each other. So for instance, the effect of one unit does not lead to greater or lesser effect for another unit.

Third, while not an assumption, it is important to note that randomization works asymptotically and not necessarily in the sample. So even if a treatment leads to greater scores for the treatment group than for the control group, a given study might not show this effect because of unhappy randomization: Those assigned to the treatment condition are systematically different from those in the control condition. Given randomization, such differences are unlikely, but they are not impossible.

Observational Studies

Studies without random assignment are called *observational* studies in the RHCM. Here we have a treatment variable and a host of pretreatment variables, often called *covariates*, and the outcome or dependent variable. We have sets of units measured under different levels of the treatment variable. The issue is how to use the covariates to make adjustments for these covariates to obtain a measure of the causal effect of the treatment across units. The solution within the RHCM is something called *propensity scores* (Rosenbaum, 1984; Rosenbaum & Rubin, 1983; 1984).

A propensity score is the probability that a given unit would be assigned to the treatment group. Although there are some differences, propensity score analysis basically proceeds as follows:

Step 1: Treat the treatment variable as an outcome variable in logistic regression analysis. Use the set of covariates to predict it.

Step 2: From Step 1, compute the propensity score as the probability of a person being in the treatment group. Create about 5 groups of about equal size for whichever group (treatment or control) is smaller in number and call this variable a *classification variable*.

Step 3: Examine the effect of the treatment on the outcome controlling for the effect of the classification variable that is treated as a nominal variable.

In essence, propensity scores match control and treated units on a set of covariates.

The key assumption of propensity score analysis is the covariates that are used to create it exhaust the possible set of covariates that should be used, an assumption called *strong ignorability*. The assumption is that "all variables related to the *both* outcomes and treatment assignment are included in" the analysis (Rosenbaum & Rubin, 1984). Another way of saying this is that there are no omitted or spurious variables.

Treatment Compliance Studies

The utility of the RHCM is very apparent in the analysis of noncompliance of clinical trials. Here we only briefly outline the approach, and we refer the reader to more extended discussions (Imbens & Rubin, 1997). The problem is that there are two randomly formed treatment groups, but some treated units opt not to receive the treatment (i.e., and so become control group members) and some controls end up receiving the treatment. There is then non-compliance.

With the RHCM, we ask the following questions: Would not some of the treated units ended up treated if they were assigned to control group (i.e., always takers)? Would not some of the controls ended up not complying if they were assigned to the treatment group (never takers)? Imbens and Rubin (1997) derive an estimator of the causal effect of the treatment that turns out to be the same an instrumental variable estimator (where manipulated treatment serves as the instrumental variable). Given the assumption that there are no defiers (people who would always opt for the treatment that they are not a member of), the estimator is a valid estimate of the causal effect, at least of the compliers.

Other Possible Implementations

Accumulated experience in psychological research suggests that there are several other possible ways, outside of the implementations discussed in RHCM, in which causal effects can be estimated. In the following paragraphs we discuss some of these.

Repeated Measures Revisited

The first additional implementation that we would suggest permits causal inference is a variation on repeated measures implementation but without the implausible temporal stability assumption. We to assume, as before, causal transience; that is, that the effect of the treatment variable is transient, affecting current responses only, but quickly dissipating and showing no subsequent effects when further measures of the response variable are taken on the unit.

In this implementation, there are multiple units exposed in sequence to the two levels of the manipulated treatment variable. However, these units are exposed to those two levels in different sequences. For some of them $\boldsymbol{Y_c(u)}$ is measured first and subsequently $\boldsymbol{Y_t(u)}$ is measured. For others, $\boldsymbol{Y_t(u)}$ is measured first and subsequently $\boldsymbol{Y_c(u)}$. Without the assumption of temporal stability, units are assumed to change over time, either because of maturation or history or some other effects. For individual units, this means that the second measurement of $\boldsymbol{Y}$ will differ from the first, not because of any change in the independent variable, but because of spontaneously occurring changes in the response over time. What is necessary, however, is that those changes over time in responses need to be unconfounded with whether units are first measured in $\boldsymbol{t}$ and subsequently measured in $\boldsymbol{c}$ or whether they are first measured in $\boldsymbol{c}$ and subsequently measured in $\boldsymbol{t}$. The way to accomplish this is to use the randomization of units implementation, but instead of randomizing units to the different treatment levels, units are randomized with respect to sequence. If this is the case, then on average the spontaneous induced changes in responses over time, induced by maturation and history and other factors, are on average unconfounded with the treatment differences. One then compares the two average responses of all units, their response in $\boldsymbol{t}$ and their response in $\boldsymbol{c}$, and from these two averages estimates the treatment effect.

Regression Discontinuity

In 1977 Rubin published a paper entitled "Assignment of treatment group on the basis of a covariate." In that paper, while he set out the basics of what subsequently became known as RHCM, he considered the regression discontinuity design, in which assignment to one of two treatment conditions is based on some measured variable, *X*. He considered both the classic regression discontinuity design, were there is a fixed cutoff score on *X* and units below that are assigned to one level of the treatment variable and units above that score are assigned to the other level, and a design where *X* is used probabilistically to assign units to the treatment variable. At any particular value of *X*, unit are randomly assigned to one or the other treatment level, and the probability that they are assigned to one level, as opposed to the other, increases as *X* increases in value.

As numerous authors have since argued, causal inferences about the treatment effect can be made in this design so long as one adjusts the response variable for *X*. The problem is that in order to do the appropriate adjustment, one must make assumptions about the functional form of the *X*-*Y* relationship, e.g., whether it is a linear or some other relationship. If the functional form is misspecified, then the estimate treatment effect can be biased.

Given the emphasis of the RHCM on comparison, this design more comfortably fits within the RHCM when the assignment rule is probabilistically based on *X*, so that at any given value on *X* there are some units that are in one level of the treatment variable and others who are in the other level. In this case, and given random assignment within those levels of *X* (albeit with unequal probabilities), one can then compare the scores on the response variable of units at particular *X* levels, some of whom are in one level of the treatment variable and others in the other level. This then becomes a design that operates much like a randomized experimental one, except that the treatment comparisons occur within levels of the *X* variable.

When treatment assignment is based on *X* exactly, with a fixed cut-off score, rather than probabilistically, it is difficult to think about equivalent observations in the two treatment groups, equivalent on the *X* variable, whose *Y* response scores could be compared. By definition there exist no units in the two treatment conditions who are equivalent on *X*, because in fact there is a fixed cut-off that determines who is in which treatment condition. From this point of view, it seems that the fixed-cutoff regression discontinuity design is difficult to fit easily under the rubric of RHCM. Nevertheless, most discussions of this design consider it to have very high internal validity, so long as the correct *X*-*Y* functional form has been specified.[2]

Over-time Models without Random Assignment

One of the concerns that we have with the RHCM is that it does not have a category for what have been called *quasi-experiments*. Instead, RHCM seems to only have a simple dichotomy between randomized studies and observational studies. However, for over 40 years quasi-experiments (Campbell & Stanley, 1963; Cook & Campbell, 1979; Judd & Kenny, 1981; Shadish et al., 2002) have been thought to offer a bridge between the two types of studies.

Quasi-experiments are like observational studies in that units (or times) are not randomly assigned to conditions. However, they approximate experiments because they model the effects of time. We consider in this section the two major types of quasi-experiments: the nonequivalent control group design and the interrupted time series design.

Nonequivalent control group design. In the simplest and most common version of this design, all units are pretested and then a nonrandom subset receive the treatment and the rest serve as controls. This design is the prototypical quasi-experimental design.

In the quasi-experimental tradition, the pretest is not just one covariate of many covariates. Rather, it is a very special variable because it indexes the degree to which there are differences between the two groups on the outcome variable that are due to selection. Thus, the pretest can be used to quantify the amount of selection in the study.

One strategy for the analysis of the nonequivalent control group design is a propensity score analysis that treats the pretest instead of the posttest as the outcome variable. If there were treatment effects in such an analysis, then we might worry that non-ignorable variables are not included. Thus, a key question when this design is used is whether there are any pretest differences once propensity scores have been controlled.

Typically we will be unable to explain why there are any such pretest differences. Thus, we would conclude that there are hidden or latent variables that we need to be controlled. How can we control for the effects of these hidden variables, because they have not been measured? Here is where having a quasi-experimental design helps us. We develop a model of change and that model allows us to forecast what will happen to the pretest difference between experimental and control group over time.

Following Judd and Kenny (1981), there are two major possibilities: regression adjustment and change score analysis. In regression adjustment, we presume that the gap seen at the pretest attenuate over time, akin to the change expected due to regression toward the mean. Alternatively, in change score analysis we presume that the gap persists over time. Thus, if there is a five-unit difference in the pretest, the expectation is that the same difference will emerge at the posttest if there were no treatment differences.

As elaborated in Judd and Kenny (1981) and Kenny (1975), we assume that the hidden or latent variable is correlated with the pretest true score and that the hidden variable, called the assignment variable in Judd and Kenny (1981), has no effect on the outcome that is not mediated by the pretest true score. Because it is the pretest true score, not the pretest itself, that is the mediator, allowances for measurement error in the pretest must be made.

A different set of assumptions can be made about the hidden variable. We might alternatively assume that it has the same effect on the pretest and posttest. Such an assumption implies that the gap between groups seen at the pretest would persist over time if there were no treatment effect. The proper analysis of the data then becomes raw change score analysis.

Change score analysis can be recast as a growth-curve model (Raudenbush & Bryk, 2002). If we adopt a change score approach, we are presuming that growth is linear. Given that the time between pretest and posttest is the same for all units, a growth curve model of linear growth would permit the use of raw change score analysis. However, if the time interval between pretest and posttest varies, then growth-curve analysis can model those differences and so, given the assumption of linear growth, improve on the estimates of treatment effects.

If growth is nonlinear then the modeling becomes more complicated. Because there are only two time points, it is impossible to measure any nonlinear function for any person. However, if it can be assumed that growth is nonlinear with respect to some variable, usually age, then a nonlinear growth curve can be determined if that variable is measured. So for instance, it might be assumed that growth is an exponential function of age. Then the growth for any given individual can be determined by his or her age at the pre- and posttest. In this way selection by maturation effects can be modeled appropriately.

There are special variants of this design (e.g., having a pre-pretest) and Shadish et al. (2002) detail them. Having more measurements and more outcome variables presents special opportunities for analysis.

Interrupted time-series design. In the simplest version of the design, a single unit is observed for the first n_1 occasions (an initial set of observations) and then the treatment is delivered and the unit is repeatedly observed for the subsequent n_2 occasions. A treatment effect might be indicated by higher or lower scores for treated than control observations.

As we stated elsewhere (Judd & Kenny, 1981, pp. 134-136), causal inference for this design is very tenuous. Because the assignment of occasions to treatment conditions is done non-randomly, we do not what variable brings about assignment. Particularly problematic is the threat of regression toward the mean because interventions are often introduced when scores are particularly extreme (either high or low).

There are various strategies that can increase the interval validity of the conclusions from this design. These strategies are detailed in Shadish et al. (2002) and we mention them only briefly. For example, one possibility is multiple time series with the timing of the intervention varying across the different time series. With such a design history threats to internal validity can be reduced.

Generally for the interrupted time series design, the assumption of linear growth is made. Such an assumption can be tested and nonlinear growth or change can be estimated.

One potential advantage of time-series designs is the possibility of weakening the transience assumption that RHCM has emphasized for repeated measures. A particular decay function must be assumed. But with the times-series data the decay or even growth function can be measured. McCain and McCleary (1979) provide the details.

One way of modeling data that are several times series is to use multilevel modeling. In this case we might have many respondents who are measured on many occasions. There are two levels in the analysis, one being the person and the other being the observation within person. We can then estimate the effect of the treatment on outcome for each person and we can test if there is variation between units in these effects. The reader should consult Bolger, Davis, & Rafaeli, 2003) for more details on such analysis.

Observational Studies

As we stated earlier, in observational studies there are two groups of observations, treated and control, from different units, and a set of variables, commonly called *covariates*, that are measured before the intervention was implemented. Following Cronbach, Rogosa, Floden and Price (1977) see also Reichardt (1979), there are two fundamental choices about how to use the covariate information to adjust for the bias due to selection. One is to model the assignment process and the other is to model the outcome variable. Propensity score analysis adopts the strategy of modeling assignment. The more traditional analysis is to add them as covariates to the model for the outcome variable. This analysis, which we call regression modeling, amounts to using the covariate variables to model the outcome. In thinking about the choice between these two approaches, we wish to make several points.

Both strategies will fail if there are any omitted variables, which are in RHCM parlance *nonignorable* variables. While sensitivity analyses are possible, there fundamentally is no way to know for sure whether all the variables that are needed are in either model.

There are some attractive features to propensity score analysis. It can take a large number of variables and reduce all of their effects to a single variable. Thus, there may be little loss of degrees of freedom. Also in the construction of the propensity score, the treatment effects are not estimated, whereas in traditional regression modeling the treatment effect is estimated. So for the regression modeling there may be both conscious and unconscious temptations to cheat; i.e., to include or not include a covariate in the analysis based on how it affects the treatment effect estimate. There is, however, one very serious disadvantage in using propensity score analysis. There may be a variable that is critical to assignment, but that variable has absolutely no effect on the outcome. If so the inclusion of such a variable in the construction of the propensity score seriously lowers the precision of the estimate of the treatment effect. Perhaps there might be some way to modify propensity score analysis so that variables that uniquely predict the outcome are more likely to be included, whereas variables that do not are less likely to be included.

We wish to make one final point about propensity score analysis. As has been explained, this strategy attempts to model the selection or assignment process to treatments. There are usually two classes of variables used in this modeling. The first set consists of the demographic variables of the units (e.g., age, gender, and ethnicity). The second set consists of psychological variables (e.g., motivation to enter treatment). It seems likely that there would be measurement error in such measures and allowances for that measurement error should be undertaken.

Alternatives to the RHCM

We need to be clear. We think, with some modifications, the RHCM presents a reasonable way to define causation. However, we do think though that by starting with experimentation, it tends to restrict the thinking of researchers about causation. Statements like "No causation without manipulation" (Rubin, 1986, p. 962) over-emphasize the importance of experimental manipulation in the inference of causal effects. There are alternatives that deserve consideration.

The major competitor, as we see it, is the Suppes (1970) model of probabilistic causation. It states that there are three necessary conditions to causation: temporal precedence, association, and non-spuriousness. Holland (1986) is critical of this model because it is conditional in the sense that any causal effect can eventually be revealed to be biased due to eventually finding a spurious explanation.

The more contemporary realization of the Suppes model is structural equation modeling (SEM). Pearl (2000) argues that SEM and RHCM (what he calls Neyman-Rubin potential-outcome model) are in fact "mathematically equivalent" (p. 134). Pearl (2000) goes onto say that the RHCM "has only been partially formalized and

… rests on an esoteric and seemingly metaphysical vocabulary of counterfactual variables that bears no apparent relationship to ordinary understanding of cause-effect processes." For this reason we prefer the SEM approach over the RHCM, especially when one is considering nonexperimental studies. It makes the researcher focus on the effects of variables that need to be controlled.

We see as the major drawback of the RHCM as not the model itself, but rather the rather limited set of options that are usually given for being able to estimate causal effects. As we have tried to show, there are more options than just randomization, repeated measures (with very restrictive assumptions), and propensity score analysis.

Conclusion

There is a set of principles about causal inference that we believe would likely elicit widespread agreement:

1. All causal inference involves assumptions, the set of which can be referred to as a *model*.
2. The assumptions of any given model cannot be known with certainty to be true. Some will necessarily remain untestable.
3. Thus, the strength of causal inference (i.e., internal validity) rests on the relative, not absolute, plausibility of a model's assumptions.

Regardless of the model of causality, these principles hold true.

We often see variants of the following statement: "Because the study is not a randomized experiment, we cannot make any causal inferences." We strongly disagree with this statement. Certainly the causal inferences from non-experimental research are weaker than that from experimental research (although if the regression discontinuity design is considered non-experimental its internal validity very nearly approximates that off an experiment). We can draw causal inferences from correlational studies, albeit often very weak ones.

To illustrate the point, consider the following two research studies:

A: Smith administers an injection of drug A to 10 self-selected HIV+ patients (no control group) and six months later in none of the 10 is there any evidence of the virus.

B: In a double blind study, Jones administers an injection of drug B to 200 randomly selected HIV+ patients, half of whom receive a placebo; six months later there is a statistical reliable 5% reduction of viral load in those who received the drug.

Must we conclude only that Jones established a causal relationship between drug and viral load and that Smith has some very preliminary data? If you or someone you cared about were HIV+, would you use Smith's drug or Jones' drug? The answer seems obvious to us. Other factors besides the design features that we have emphasized are relevant for determining causality [see for example Abelson's (1995) MAGIC]. As the example illustrates, the absolute magnitude of an effect is one such piece of information.

The purpose of this paper is not encourage sloppy or poorly designed studies. The purpose rather is to argue that causal analysis is a fundamentally human enterprise that cannot be routinized. While randomized experiments are one key tool in establishing causality, they are only one tool. Other, albeit weaker, tools are available. The assumptions underlying any model of causality from any research design need to be appropriately examined and questioned. There is not a single experimental path to causal inference.

References

Abelson, R. P. (1995). *Statistics as principled argument*. Hillsdale, NJ: Erlbaum.

Bolger, N., Davis, A., & Rafaeli, E. (2003). Diary methods: Capturing life as it is lived. *Annual Review of Psychology*, in press.

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching*. Chicago: Rand-McNally.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation*. Chicago: Rand-McNally.

Cronbach, L. J., Rogosa, D. R., Floden, R. E., & Price, G. G. (1977). *Analysis of covariance in nonrandomized experiments: Parameters affecting bias*. (Occasional paper.) Berkeley, CA: Stanford University, Stanford Evaluation Consortium.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica, 47,* 153-161.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*, 945-960.

Holland, P. W. (1988). Causal inference, path analysis, and recursive structural equation models (with discussion). In C. Clogg (Ed.). *Sociological Methodology 1988* (pp. 449-493). Washington, D.C.: American Sociological Association.

Imbens, G. W., & Rubin, D. B. (1997). Bayesian inference for causal effects in randomized experiments with noncompliance. *The Annals of Statistics, 25*, 305-327.

Judd, C. M., & Kenny, D. A. (1981). *Estimating the effects of social interventions*. New York: Cambridge.

Kenny, D. A. (1975). A quasi-experimental approach to assessing treatment effects in the non-equivalent control group design. *Psychological Bulletin, 82*, 345-362.

McCain, L. J., & McCleary, R. (1979). The statistical analysis of the simple interrupted time-series experiment. In T. D. Cook & D. T. Campbell, *Quasi-experimentation: Design and analysis for field settings* (pp. 233293). Boston: Houghton Mifflin.

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

Raudenbush, S. W., & Bryk, A. S. (2002). Hierarchical linear models: Applications and data analysis methods (2nd. Ed.). Thousand Oaks, CA: Sage.

Reichardt, C. S. (1979). The statistical analysis of the nonequivalent group designs. In T. D. Cook & D. T. Campbell, *Quasi-experimentation: Design and analysis for field settings* (pp. 147-205). Boston: Houghton Mifflin.

Reichardt, C. S. (2003). *On the logic of estimating treatment effects*. Unpublished paper, University of Denver.

Rosenbaum, P. R. (1984). From association to causation in observational studies: The role of tests of strongly ignorable treatment assignment. *Journal of the American Statistical Association, 79*, 41-48.

Rosenbaum, P. R., & Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*, 41-55.

Rosenbaum, P. R., & Rubin, D. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association, 79*, 516-524.

Rubin, D. B. (1977). Assignment of treatment group on the basis of a covariate. *Journal of Educational Statistics, 2*, 1-26.

Rubin, D. B. (1980). Discussion of "Randomization analysis of experimental data: The Fisher randomization test." *Journal of the American Statistical Association, 75*, 591-593.

Rubin, D. B. (1986). What if's have causal answers. *Journal of the American Statistical Association, 81*, 961-962.

Rubin, D. B. (1990). Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science, 5*, 472-480.

Rubin, D. B. (1991). Practical implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism. *Biometrics, 47*, 1213-1234.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference. Boston, MA: Houghton Mifflin Company.

Suppes, P. (1970). *A probabilistic theory of causality*. Amsterdam: North-Holland.

West, S. G., Biesanz, J. C., & Pitts, S. C. (2000). Causal inference and generalization in field settings: Experimental and quasi-experimental designs. In H. T. Reis & C. M. Judd, (Eds.), *Handbook of research methods in social and personality psychology* (pp. 40-84). New York: Cambridge.

Footnotes

[1] It is important to note that Rubin, in a commentary to the Holland (1986) paper, suggests that he may not see this ***A/ S*** distinction in the same way. Rubin (1986), for instance, does acknowledge that recent surgical and hormonal possibilities make gender potentially manipulable. However, he does endorse the general principle that unless a variable can be manipulated, its causal effect cannot be assessed.

[2] We do not consider in this paper selection models such as those developed by Heckman (1979) and others. These models are like the regression discontinuity design in that they make an assumption (e.g., selection from a normal distribution instead of selection using a variable that has linear relationship to outcome) to identify a causal effect.