

Choosing an Identifying Set of Matching or Conditioning Variables *

Adam N. Glynn[†] Kevin M. Quinn[‡]

June 6, 2008

Abstract

Political scientists estimate average causal effects with regression or matching techniques, but both techniques require the user to choose a set of matching or conditioning variables. In this paper, we show that the standard advice from both frameworks on how to choose an identifying set of variables is often insufficient and at times misleading. Furthermore, we argue that structural causal models (SCMs) provide an attractive framework to reason about this problem. SCMs provide simple rules for choosing an identifying set of variables given particular causal assumptions. This framework is equivalent to the Neyman-Rubin framework and common adjustment methods—such as those based on generalized linear models—can be analyzed within the framework. Finally, we demonstrate that SCMs allow the specification of causal modeling assumptions in a manner that is compatible with the mechanistic view of causation commonly invoked by political scientists.

*The authors thank Thomas Richardson for introducing them to literature of graphical causal models, and also thank Neal Beck, Andy Eggers, Jim Greiner, Ben Hansen, Gary King, Judea Pearl, Anton Westveld, two anonymous referees, and the participants of 2007 Summer meetings of the Society of Political Methodology for their helpful comments and suggestions. The usual caveat applies. In addition, Quinn thanks the National Science Foundation (grants SES 03-50613 and BCS 05-27513) and the Center for Advanced Study in the Behavioral Sciences for its hospitality and support.

[†]Department of Government and The Institute for Quantitative Social Sciences Harvard University, 1737 Cambridge Street, Cambridge, MA 02138. aglynn@iq.harvard.edu

[‡]Department of Government and The Institute for Quantitative Social Sciences Harvard University, 1737 Cambridge Street, Cambridge, MA 02138. kevin.quinn@harvard.edu

1 Introduction

The choice of what variables to adjust for when attempting to make causal inferences from observational data is one of the most important decisions made by applied researchers. Unfortunately, it has also received less scholarly attention than the question of how to perform an adjustment *given* the appropriate set of covariates. Misunderstandings abound.

The classical econometric advice is that when a linear model is appropriate for causal inference and one is unsure whether a measured pre-treatment¹ variable should be included or not it is best to err on the side of including the potentially irrelevant covariate. Kmenta (1986) sums this up nicely:²

The conclusion, then, is that if the specification error consists of including some irrelevant explanatory variables in the regression equation, the least squares estimators of the regression coefficients are unbiased but not efficient. The estimators of the variances are also unbiased, so that, in the absence of other complications, the usual tests of significance and confidence intervals for the regression coefficients are valid. (p. 449)

This sort of argument leads many researchers to adjust for a very large number (often upwards of 15-20) measured pre-treatment variables without thinking carefully about how these might or might not be related to the outcome, treatment assignment, and each other causally.

Some researchers take a slightly different view. In assessing whether a particular set of pre-treatment variables \mathbf{Z} acts as a confounder (i.e., failing to adjust for \mathbf{Z} would bias causal effect estimates) they tend to focus on whether various associations in the observed pre-intervention distribution are non-zero. More specifically, they judge the effect of X on Y to be confounded by \mathbf{Z} if

1. X is not statistically independent of \mathbf{Z} ; and
2. Y is not conditionally independent of \mathbf{Z} given X

¹Less careful researchers will not make the restriction to pre-treatment variables even though this is known to be problematic.

²For similar sentiments see also Fox (1997, p. 235 *fn* 50), Pindyck and Rubinfeld (1998, p. 187), and Greene (2000, p. 338)

For examples see Schlesselman (1982); Rothman (1986); Rothman and Greenland (1998) along with the discussion in Chapter 6 of Pearl (2000). Because the relevant associations are directly observable from observational data such a criterion has great appeal. The criterion can easily be checked, and if \mathbf{Z} is judged to be a confounder, then it is adjusted for in some way, e.g., via regression, stratification, matching, etc.

Much of the empirical work in leading political science journals tends to (at least implicitly) take one of these two positions when defending the choice of adjustment variables.³ Unfortunately, neither of these approaches is guaranteed to identify sets of adjustment variables that are sufficient to control confounding— *even when some subset of measured pre-treatment variables available to the researcher is sufficient to control confounding*. This is true regardless of the adjustment method. The fundamental problem with both of these approaches is that they seek either implicitly or explicitly to use characteristics of the observed data distribution *without additional causal assumptions* to determine the appropriate set of adjustment variables.

The Neyman-Rubin causal model (Neyman et al., 1935; Rubin, 1974, 1978) offers a major step forward in that it focuses a researcher’s attention on whether conditional ignorability (a causal assumption) holds for a given set of adjustment variables \mathbf{Z} . The advantage here is that if conditional ignorability does hold given \mathbf{Z} then adjustment for \mathbf{Z} is *guaranteed* to be sufficient to control confounding. The major drawback is that, by itself, the Neyman-Rubin framework provides little guidance as to what sets of background variables are likely to produce conditional ignorability.

The advice that is offered by proponents of the Neyman-Rubin model focuses on the model for treatment assignment; i.e., the explanatory variable of interest (Rubin, 2004). This is in contrast to the classic econometric treatment of regression analysis that focuses almost exclusively on the

³Clarke (2005, 2006) provide examples of political science articles that include a large number of adjustment variables seemingly for fear of omitted variable bias. He then goes on to correctly note that, within the context of generalized linear regressions, bias is not necessarily monotonically decreasing in the number of background variables adjusted for. Much of the discussion in that article is focused on whether a covariate is “relevant” which we assume means whether the covariate exerts a causal effect on the outcome variable. The results we describe in this paper are more general in that they apply to all recursive causal models— not just causal generalized linear models. Further, we will see that there are situations where adjusting for an “irrelevant” variable, i.e., a variable that exerts no causal effect on either the outcome or treatment assignment, will bias estimates of causal effects.

model for the outcome variable. To quote two leading architects of the Neyman-Rubin perspective:

...we cannot be sure that treatment assignment is strongly ignorable given the observed covariates because there may remain unmeasured covariates that affect both outcomes and treatment assignment. (Rosenbaum and Rubin, 1984, p. 522)

However, as we demonstrate in this paper, advice of this type has led to a number of widely held beliefs that are not entirely accurate. For example, *none* of the following statements are strictly true.

- Bias cannot increase as an additional pre-treatment variable is adjusted for.
- All pre-treatment variables that are associated with treatment assignment as well as with the outcome given treatment assignment are confounders and need to be adjusted for.
- Balance on all measured pre-treatment variables is necessary for consistent estimation of causal effects.
- Balance on all measured pre-treatment variables is sufficient for consistent estimation of causal effects.

These inaccurate beliefs arise because the assumption typically used to identify average causal effects—the so-called conditional ignorability of treatment assignment assumption—is an untestable⁴ *global* assumption that can, at times, be difficult to assess. The reliance on counterfactual quantities and the associated lack of testability is not specific to the Neyman-Rubin framework. However, the reliance on a single global assumption is not a property of all causal models. By replacing this single large assumption with a series of local assumptions the structural causal models (SCMs) of Pearl (1995, 2000) offer researchers additional means of assessing the adequacy of various adjustment strategies, and clarify some aspects of procedures that have come out of the standard Neyman-Rubin framework.

In this paper we advocate a hybrid approach that is consistent with the Neyman-Rubin framework but that specifies underlying mechanisms in a non-parametric fashion. More specifically, we show how SCMs add enough structure to the Neyman-Rubin model so that sufficient conditioning sets can be chosen. Unlike traditional linear structural equations models that make strong

⁴Conditional ignorability of treatment assignment refers to the assumption that the counterfactual values of the outcome variable under treatment and control are independent of the actual treatment assignment given some set of adjustment variables. Note that tests for balance are not tests for ignorability of treatment assignment. Ignorability is defined in terms of counterfactual quantities and hence cannot, by its very nature, be subjected to statistical tests.

functional form assumptions, SCMs only make assumptions about which variables determine the potential outcomes of other variables. Furthermore, the inclusion of mechanisms helps applied researchers think about the assumptions that are necessary to make causal inference. Nearly all theoretically-informed work in political science has some discussion of the causal mechanisms hypothesized to have generated the data available to researchers. Notable examples include, among others, the “funnel of causality” of the Michigan model of voting (Campbell et al., 1960), Fearon’s (1995) work on rationalist explanations of war, and Cox (1997) on the mechanisms that produce the effective number of parties. Indeed, some have argued for an explicit focus on causal mechanisms as opposed to universal laws and grand theories (Elster, 1989a,b, 1998; Collier and Brady, 2004).

This article proceeds as follows. In Section 2, we pose two stylized questions that are designed to highlight the complexity of the covariate selection problem. In Section 3 we briefly review the key components of the Neyman-Rubin causal model. In Section 4 we describe the SCM, paying special attention to points of commonality and complementarity between the Neyman-Rubin model and the Pearl model. In Section 5 we introduce directed acyclic graphs as a representation of structural causal models, discuss their probabilistic implications, and present a simple graphical criterion for choosing an identifying set of adjustment variables. In Section 6 we revisit the covariate selection problem the two examples from Section 2 and a more complicated parametric example. These examples are designed to make clear some of the problems with standard approaches to choosing adjustment variables and how graphical methods can provide additional leverage on this difficult problem. The final section concludes.

2 Stylized Questions for Researchers

The following two subsections present stylized questions that are similar to those faced by empirically oriented political scientists. Note that while these examples do not represent well-designed studies, they are not so different than those that are routinely conducted within the social sciences. In Section 6 we use the machinery developed in this paper to answer these questions,

and some readers may find it useful to skip to the solutions before reading the technical details in Sections 3, 4, and 5.

2.1 The Effect of British Colonial History on Authoritarianism

Suppose researchers are interested in the average causal effect of British colonial history on current authoritarianism as coded by Polity score. Furthermore, they are willing to assume that a linear additive regression that includes a dummy variable for British colonial history (BCH) and a recent ethnolinguistic fractionalization (ELF) index will be sufficient to approximately satisfy all of the classical econometric linear model requirements (Wooldridge, 2000; Greene, 2000). They consider fitting two regressions, one with both variables and one with only BCH, and are faced with two questions:

1. Given the current set of assumptions, do they have enough information to state that one/both of the least squares estimators is unbiased for the average causal effect of BCH on Polity score?
2. If not, do they have enough information to say which estimator will have less bias, or what additional assumptions would be necessary to remove the bias for each of the estimators?

We provide answers to these questions in Section 6.1.

2.2 The Effect of an Intervention on Inter-Ethnic Cooperation

Suppose researchers are interested in determining the effect of a novel two-week within-school program on inter-ethnic cooperative behavior. It is assumed that there are two mutually exclusive and exhaustive ethnic groups. The researchers adopt the following research design. The population of interest consists of all entering students in two large high schools. In what follows, the implicit treatment variable X is the exposure to the novel program (measured as unexposed = 0, exposed = 1).

The program is implemented in only one of the two schools just prior to the beginning of the school year to all first year entering students. The school that implements the program has a longstanding reputation as the more ethnically inclusive of the two schools under study. The schools are alike in all other relevant respects. At the beginning of the school year, each entering

student in both schools is taken to a mobile experimental laboratory where s/he is told that s/he will play the role of the first-mover in a one-shot divide-the-dollar game. Via a computer terminal, the student makes a single take-it-or-leave-it offer $y \in \{\$0.01, \$10, \$19.99\}$ to the second (unseen) player. Before playing the game, the first-mover is told that if the second player accepts the offer the first-mover will receive $\$20 - y$ and if the second player rejects the offer the first-mover will receive $\$0$. Unknown to the first-mover, the second “player” is a computer program that randomly chooses to either reject or accept the offer. Nonetheless, the first-mover is told that the second player is a member of the ethnic group different from the first-mover’s group. The goal of the study is to ascertain the effects of exposure to the novel program (X) on altruistic inter-ethnic behavior (Y).

Because the program was implemented in the school known to be more ethnically inclusive, there is concern among the researchers that the ethnocentrism of a student’s parents might exert an effect on the school the child attends (and hence exposure to the program) as well as the child’s ethnocentrism and ultimately the child’s behavior in the experimental game. However, the researchers are willing to assume that the background factors that influence the outcome variable (e.g. ability to recognize the strategic implications in a divide the dollar game) are independent of school choice and student ethnocentrism (and their respective background factors).

In the hope of developing a proxy for the parents’ ethnocentrism, the researchers administer a short survey to each student before they are exposed (or could have been exposed) to the new program. Each student’s survey responses are used to construct a measure of his/her latent ethnocentrism. This variable is labeled Z (measured as low ethnocentrism = 0, high ethnocentrism = 1). After the study has been completed, the researchers find (as they had expected) that there is a significant association between Z and X and a significant association between Z and Y within levels of X . Table 1 represents the joint probability function of X, Y , and Z .

The researchers consider two matching estimators (one that matches on nothing, and one that matches on the constructed measure of ethnocentrism) and are faced with two questions:

$Z = 0$ (low ethnocentrism)		
	$X = 0$ (not exposed)	$X = 1$ (exposed)
$Y = \$0.01$	0.1728	0.2219
$Y = \$10.00$	0.1224	0.1868
$Y = \$19.99$	0.0144	0.0220

$Z = 1$ (high ethnocentrism)		
	$X = 0$ (not exposed)	$X = 1$ (exposed)
$Y = \$0.01$	0.0772	0.0281
$Y = \$10.00$	0.1012	0.0368
$Y = \$19.99$	0.0120	0.0044

Table 1: *Joint Probability Function For Variables in Ethnic Cooperation Example.*

1. Given the current set of assumptions, do the researchers have enough information to state that one/both of the matching estimators will identify the average causal effect of the program on inter-ethnic cooperation?
2. If not, do they have enough information to say which estimator will have less bias, or what additional assumptions would be necessary to remove the bias for each of the estimators?

We provide answers to these questions in Section 6.2.

3 The Neyman-Rubin Model

In the Neyman-Rubin model, causal effects are defined in terms of potential outcomes: $Y(x, u)$ (i.e. the potential outcome in unit u if X would have been set equal to x) (see Rubin (1978), Rosenbaum and Rubin (1983), Holland (1986)). Here u is thought of as a unit-specific index, and therefore captures any individual-specific effects. It is tempting to think of units as individuals, schools, etc., but in actuality it is more accurate to think of the units as individuals, schools, etc. under a particular set of exogenous background conditions. Thus an individual at 9:00AM and the same individual at 10:00AM may very well be considered different units. If the value of X received by one unit does not affect the outcomes for other units, then given x and u , $Y(x, u)$ is completely determined. This assumption of non-interference is sometimes called SUTVA (see

Angrist et al. (1996)), and we will utilize this assumption throughout this paper.

3.1 Unit-Specific Causal Effects

In the Neyman-Rubin model, the potential outcomes are used to define unit-specific causal effects. For simplicity in presentation, we assume that X can only take on the values zero and one.⁵ Therefore, the unit-specific causal effect of $X = 1$ on Y relative to the effect of $X = 0$ in unit u is calculated by comparing $Y(1, u)$ to $Y(0, u)$. A common means of comparison is the difference::

$$Y(1, u) - Y(0, u).$$

The key idea is that if it were possible to observe $Y(1, u)$ and $Y(0, u)$ for the two levels of the treatment variable (e.g. active treatment and control), then we could observe the unit specific causal effect.

If we assume consistency (Robins, 1986) of the observed outcomes, then we may observe one of these two outcomes for each individual. This assumption requires that the observed outcome for each unit $Y(u)$ matches the potential outcome for unit u for the observed value of X . Formally, this can be written as the following:

$$\mathbf{X}(u) = x \implies \mathbf{Y}(u) = \mathbf{Y}(x, u).$$

and if this holds, then our binary treatment example, the observed Y can be written as the following:

$$Y^{obs}(u) = X(u) \cdot Y(1, u) + (1 - X(u)) \cdot Y(0, u)$$

Unfortunately, consistency does not allow the unit-specific causal effect to be directly observed since u only gets one of either $X = 0$ or $X = 1$ but never both. Holland (1986) calls this the fundamental problem of causal inference.

3.2 Population Causal Effects

Given the impossibility of observing individual causal effects, inference is usually confined to the characteristics of populations (sometimes the observed sample of individuals is taken as the

⁵The extension to polytomous or continuous treatment variables will not complicate the discussion of identification in this paper, but will complicate the choice of adjustment strategy.

entire population). For simplicity, we assume throughout this paper that the parameter of interest is the *average causal effect* from $X = 0$ to $X = 1$. This is defined as

$$ACE \equiv \mathbb{E}[Y(1) - Y(0)],$$

where the expectation merely represents an average over the units (or the distribution of pertinent background factors) in the population of interest. This parameter has a number of useful properties including the usual decomposition of the expectation of sums which allows us to separately consider the average potential outcomes under treatment and control.

$$ACE \equiv \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$$

Unfortunately, these averages are not observed in general. Instead we observe averages of potential outcomes over the subpopulations that actually received treatment and control. Hence we can identify the potentially similar parameter that Holland (1986) calls the *prima facie* average causal effect (the second line is due to consistency):

$$\begin{aligned} ACE^{pf} &\equiv \mathbb{E}[Y(1)|X = 1] - \mathbb{E}[Y(0)|X = 0] \\ &= \mathbb{E}[Y|X = 1] - \mathbb{E}[Y|X = 0] \end{aligned}$$

3.3 Ignorability and the Identification of Population Causal Effects

The Neyman-Rubin model makes clear that the following will not hold in general,

$$\mathbb{E}[Y(0)] = \mathbb{E}[Y(0)|X = 0] \tag{1}$$

$$\mathbb{E}[Y(1)] = \mathbb{E}[Y(1)|X = 1], \tag{2}$$

because averages over subpopulations need not match averages over the population. However, it is sufficient to assume the equalities in (1) and (2) in order to identify the ACE. This assumption, sometimes known as *mean ignorability*⁶, is usually hard to justify because the subpopulation that

⁶Although we focus on identification in this paper, there are other inferential goals, and hence it is often necessary to make stronger ignorability assumptions. Rosenbaum and Rubin (1983) describes sufficient ignorability assumptions for a variety of inferential tasks.

receives treatment is often quite different from the subpopulation that receives control. Random treatment assignment for a large population is an example where the subpopulations will be similar.

It is often possible to “weaken” ignorability assumptions by conditioning on a set of background variables which we will denote as \mathbf{Z} . Hence, even if (1) and (2) do not hold, we may believe that,

$$\mathbb{E}[Y(0)|\mathbf{Z}] = \mathbb{E}[Y(0)|X = 0, \mathbf{Z}] \tag{3}$$

$$\mathbb{E}[Y(1)|\mathbf{Z}] = \mathbb{E}[Y(1)|X = 1, \mathbf{Z}], \tag{4}$$

hold for some set (or sets) of \mathbf{Z} . The equalities in (1) and (2) allow the identification of average causal effects within the strata defined by \mathbf{Z} , and these can then be combined through a weighted average to identify the overall ACE. When \mathbf{Z} lives in a high dimensional space, this averaging can present considerable practical difficulty, so in order to confine the discussion to the issues considered in this paper, we assume throughout that \mathbf{Z} is discrete and has low dimension or that the joint distribution of all variables has a simple parametric form.

4 The Structural Causal Model

The structural causal model (Pearl, 1995, 2000) and its close relatives (Spirtes et al., 1993; Robins, 1986) provide additional structure to the Neyman-Rubin model. In what follows, we adapt Pearl’s presentation to the case considered in this paper (a single intervention variable, a single outcome variable, and no interference).

4.1 A Unit-Level Causal Model

Adapting the definition from Pearl (2000), we define a unit-specific causal model for a single outcome and intervention variable to be the triple:

Definition 1 (Unit-Level Causal Model)

$$M = \langle \mathbf{U}, \mathbf{V} \equiv \{Y, X, \mathbf{Z}\}, \mathbf{h} \rangle$$

where:

1. \mathbf{U} is a set of exogenous background variables.

2. $\mathbf{V} \equiv \{Y, X, \mathbf{Z}\}$ is a set of endogenous variables. Y is the outcome variable, X is the intervention variable, and \mathbf{Z} is a set of potential control variables (some possibly unobserved),
3. \mathbf{h} is a set of functions that defines the endogenous variables (one for each endogenous variable).

Furthermore, we will utilize the following conventions/assumptions/clarifications (some portions of these statements are redundant, but we include them all in order to provide intuition):

1. The set of exogenous variables is rich enough to define a unit. Therefore, we can think of the unit index u from Section 3 as a function of the realized vector of exogenous variables $\mathbf{U} = \mathbf{u}$.
2. We assume a causal order to all endogenous variables such that an endogenous variable may not be an input to the functions of any of the endogenous inputs to its function. This type of causal model is sometimes known as recursive.
3. We assume that given the values of the exogenous variables ($\mathbf{U} = \mathbf{u}$), the endogenous variables are uniquely determined by the functions \mathbf{h} .

Given the definition and assumptions, consider the following simple example:

$$Z \leftarrow h_Z(u_1)$$

$$X \leftarrow h_X(z, u_2)$$

$$Y \leftarrow h_Y(x, z, u_3)$$

Z is a *deterministic* function of u_1 , X is *deterministic* function of z and u_2 , and Y is a *deterministic* function of x , z , and u_3 . The assignment notation (\leftarrow) above is to make clear that the assignment in these functions is asymmetric, and we label the entire model M .

The model M is non-parametric in that no assumptions are made about h_Z, h_X, h_Y, U_1, U_2 and U_3 . Given the causal order assumption, the endogenous variables can be written as a function of the exogenous variables. For example, in our simple model, y can be written as a function of all other variables and functions:

$$Y \leftarrow h_Y(h_X(h_Z(u_1), u_2), h_Z(u_1), u_3).$$

Therefore, $Y(\mathbf{u})$ denotes the unique value of Y generated by model M given $\mathbf{U} = \mathbf{u}$, and is analogous to $Y^{obs}(u)$ from the Neyman-Rubin model.

Now consider intervening in the system to set variable X equal to a particular value x without *directly* disturbing any of the other variables in the system. This involves creating a submodel in which the function for X is removed and X becomes an exogenous variable.

Definition 2 (Unit-Level Submodel) *Let M be a unit-level causal model, X be the intervention variable in the set of endogenous variables \mathbf{V} and x be a particular realization of \mathbf{X} . A unit-level submodel M_x of M is the unit-level causal model*

$$M_x = \langle \mathbf{U}, \mathbf{V}, \mathbf{h}_x \rangle$$

where

$$\mathbf{h}_x = \mathbf{h}/h_X$$

For example, consider the simple example again with X set to zero.

$$Z \leftarrow h_Z(u_1)$$

$$X \leftarrow 0$$

$$Y \leftarrow h_Y(0, z, u_3)$$

We call this system of equations submodel M_x .

Within a submodel, we can define potential outcomes that are analogous to the potential outcomes from the Neyman-Rubin model by solving for the functional output of Y under the submodel.

Definition 3 (Potential Outcome) *Let M denote a unit-level causal model, and set the intervention variable X equal to value x , then the potential outcome $Y(x, \mathbf{u})$ denotes the unique solution for Y as determined by x , the exogenous variables in the model and the functions \mathbf{h}_x .*

If u were observed, the model M and submodel M_x would define unit-specific causal effects analogous to unit-specific causal effects in the Neyman-Rubin model. Consider the following simple non-parametric example:

Suppose that a “get out the vote” (GOTV) study randomly assigns a small and geographically separated group of registered voters⁷ to either treatment (phone call) or control (no phone call), and after the election voters are observed to have voted or not voted. Unbeknownst to the study

⁷The registered voters are selected for the study so that the assumption of no interference is plausible.

designers, some of the registered voters will be so put off by a phone call that they will not vote (even if they would have voted without the phone call). If we let $X = \{0 \text{ (no call)}, 1 \text{ (call)}\}$ be treatment assignment, $Y = \{0 \text{ (no vote)}, 1 \text{ (vote)}\}$ be voting status, U_1 be the exogenous treatment randomization mechanism, and U_2 be the exogenous variable that describes each registered voter’s potential response to treatment, then this scenario can be conceptualized within the SCM framework with the following set of functional assignments:

$$x \leftarrow h_X(u_1)$$

$$y \leftarrow h_Y(x, u_2)$$

In the parlance of the Neyman-Rubin model, h_X represents the treatment assignment mechanism. Furthermore, with binary treatment and binary outcome, the domain of U_2 (and hence registered voters) can be partitioned into four potential outcome equivalence classes: those who would vote regardless of whether they received treatment or control (Always Vote: $Y(0, \mathbf{u}) = 1, Y(1, \mathbf{u}) = 1$), those who would not vote regardless of whether they received treatment or control (Never Vote: $Y(0, \mathbf{u}) = 0, Y(1, \mathbf{u}) = 0$), those who are treatable and would vote with a phone call and would not vote without (Encouraged by Call): $Y(0, \mathbf{u}) = 0, Y(1, \mathbf{u}) = 1$, and those with the potential to be put off by the phone call so that they would not vote with the call, and would vote without the call (Discouraged by Call): $Y(0, \mathbf{u}) = 1, Y(1, \mathbf{u}) = 0$. Therefore, given the values of the exogenous variables, the observed variables and potential outcomes are determined. Furthermore, unit specific causal effects are defined by the potential outcome equivalence classes. The unit-specific effect is zero for the “Always Vote” and “Never Vote” units, one for the “Encouraged” units, and negative one for the “Discouraged units”. This example demonstrates the fully nonparametric nature of the model (no additivity, monotonicity, or functional form assumptions).

In this example as in most cases, \mathbf{u} is not observed, and hence we do not observe the unit-specific causal effects. The solution, as in the Neyman-Rubin framework, is to shift inferential focus to population causal effects.

4.2 A Population Causal Model

We can create a population causal model from the deterministic causal model of the previous subsection by assuming a distribution over U .

Definition 4 (A Population Level Causal Model) *A population-level model is a pair*

$$\langle M, F_{\mathbf{U}}(\mathbf{u}) \rangle$$

where M is the unit-level causal model and $F_{\mathbf{U}}(\mathbf{u})$ is a cumulative distribution function defined over the domain of U .

Again, we note the following conventions/assumptions/clarifications:

1. We configure the exogenous variables to be independent of each other, so the distribution $F_{\mathbf{U}}(\mathbf{u})$ factors accordingly. This can be accomplished by combining dependent exogenous variables into a single exogenous variable.
2. Some authors include as endogenous all variables that are inputs to more than one function (i.e. common cause variables). We do not use this convention.
3. The distribution over the endogenous variables is uniquely determined by $F_{\mathbf{U}}(\mathbf{u})$ and \mathbf{h} .

The model M along with an assumption as to the distribution of \mathbf{U} generates the “pre-intervention” distribution $F_{\mathbf{U},\mathbf{V}}(\mathbf{U}, \mathbf{V})$, and given the assumption of a recursive causal model, this joint distribution is uniquely defined. *One can estimate the marginal distribution of the observed variables in \mathbf{V} directly from observational data without making untestable assumptions.* For example, the “pre-intervention” outcome distribution can be derived by integrating over $F_{\mathbf{U},\mathbf{V}}(\mathbf{U}, \mathbf{V})$ and is written in the standard notation $F_Y(y)$.

Continuing the GOTV example from the previous section, the finite number of treatment assignments and potential outcome equivalence classes allows the interpretation of $F_{\mathbf{U}}(\mathbf{u})$ in terms of the population proportions of individuals. Hence, $F_{\mathbf{U}}(\mathbf{u})$ describes the proportions of (Always Vote, Never Vote, Encouraged by Call, Discouraged by Call) individuals in the population, and the proportions of (Treatment, Control) individuals in the population. Furthermore, these distributions

in combination with the causal model define the proportions of (Vote, No Vote) individuals in the population.⁸

Because $F_{\mathbf{U}}(\mathbf{u})$ remains unchanged for the submodel M_x , we can ask what the probability distribution of $Y(x, \mathbf{u})$ is for a \mathbf{u} randomly drawn from the population distribution of \mathbf{U} . In other words, what is the distribution of Y in the population after the intervention on X . This quantity is denoted $F_{Y(x)}(y)$ and is called the post-intervention distribution of Y . *Post-intervention distributions are not directly estimable from observational data without untestable causal assumptions.* With probability distributions defined over post-intervention distributions, the expectations and average causal effects are written as

$$\mathbb{E}[Y(x)] \equiv \int y dF_{Y(x)},$$

and obviously

$$ACE \equiv \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$$

As in the Neyman-Rubin model, we would like to establish situations in which the observable pre-intervention distribution identifies averages over the unobserved post-intervention distribution. This task is simplified by representing SCMs as directed acyclic graphs.

5 A Graphical Criterion for Causal Identification

5.1 Directed Acycyclic Graphs (DAGs)

We begin with some basic terminology. A directed graph is defined as the following:

Definition 5 (Directed Graph) *A directed graph \mathcal{G} is a pair $\langle \mathcal{V}, \mathcal{E} \rangle$ where \mathcal{V} is a finite set of vertices (a.k.a. nodes) and \mathcal{E} is the set of directed edges (a.k.a. directed arcs or directed links).*

⁸Note that given our conventions, any dependencies between the exogenous \mathbf{u} variables need to be modeled explicitly through the redefinition of the \mathbf{U} variables. In this example, if we didn't have random treatment assignment, then we might believe that experienced GOTV workers could target their efforts at those "Encouraged" registered voters. Therefore, dependence between U_1 and U_2 could be accommodated by combining them into a single exogenous variable:

$$\begin{aligned} \mathbf{U} &\equiv \{U_1, U_2\} \\ x &\leftarrow h_X(\mathbf{u}) \\ y &\leftarrow h_Y(x, \mathbf{u}). \end{aligned}$$

Each directed edge in \mathcal{E} is an ordered pair of distinct vertices from $\mathcal{V} \times \mathcal{V}$. A directed edge $(V_i, V_j) \in \mathcal{E}$ is also denoted $V_i \rightarrow V_j$.

In this paper, we think of each $V \in \mathcal{V}$ as being a (possibly non-scalar) random variable and each directed edge $(V_i, V_j) \in \mathcal{E}$ as a causal relationship between V_i and V_j (to be made explicit later).

Within this framework, a number of additional definitions will be useful.

Definition 6 (Path) A path from V_i to V_j in a graph $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$ is a sequence of distinct nodes $V_i = X_0, \dots, X_n = V_j$ such that $(X_{k-1}, X_k) \in \mathcal{E}$ or $(X_k, X_{k-1}) \in \mathcal{E}$ for each $k = 1, \dots, n$.

Note that a path cannot visit the same node twice, and the direction of the edges does not matter. For instance, a path from V_1 to V_3 exists in each of the following four graphs.

$$V_1 \rightarrow V_2 \rightarrow V_3 \tag{5}$$

$$V_1 \leftarrow V_2 \rightarrow V_3 \tag{6}$$

$$V_1 \rightarrow V_2 \leftarrow V_3 \tag{7}$$

$$V_1 \leftarrow V_2 \leftarrow V_3 \tag{8}$$

While all of these relationships represent paths from V_1 to V_3 , it is useful to make some distinctions between these types of paths and vertices.

Definition 7 (Directed Path) A directed path from V_i to V_j in a graph $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$ is a sequence of distinct nodes $V_i = X_0, \dots, X_n = V_j$ such that $(X_{k-1}, X_k) \in \mathcal{E}$ and $(X_k, X_{k-1}) \notin \mathcal{E}$ for each $k = 1, \dots, n$. We write $V_i \rightsquigarrow V_j$ to denote a directed path from V_i to V_j .

Informally, a *directed path* from V_1 to V_k requires that all edges point toward V_k along the path. For instance, (5) depicts a directed path from V_1 to V_3 . We can distinguish further between the two-edge paths depicted in (5),(6),(7), and (8). In particular, we will call (5) and (8) chain structures, we will call (6) a fork structure, and we will call (7) a collider structure. Notice that this terminology will extend to situations where these structures appear as subpaths in longer paths. We will also find it useful to define some familial relations within directed graphs. The notions of children, and descendants can be defined as the following:

Definition 8 (Children) In a directed graph $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$ the set of children of a node $V \in \mathcal{V}$ is defined to be:

$$ch(V) = \{Z \in \mathcal{V} : (V, Z) \in \mathcal{E}\}$$

Definition 9 (Descendants) In a directed graph $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$ the set of descendants of a node $V \in \mathcal{V}$ is defined to be:

$$de(V) = \{Z \in \mathcal{V} : V \rightsquigarrow Z\}$$

Put more informally, the children of V are all nodes to which there is a directed edge from V , and the set of descendants of V consists of the vertices to which there exists a directed path from V . Analogous definitions can be constructed for parents and ancestors. Using these familial notions, we can now define the class of graphs that we will utilize throughout the rest of this paper. A directed graph that does not have cycles (i.e., no vertex in the graph is a descendent of itself) is said to be a *directed acyclic graph (DAG)*.

5.2 Representing SCMs as Directed Acyclic Graphs (DAGs)

DAGs are useful in causal modeling, because they form a compact representation of the assumptions implicit in recursive SCMs. Furthermore, a causal markov condition connects SCMs to graphical rules for deriving conditional independence relations between the pre-intervention and *post-intervention* random variables. Therefore, graphs provide a shorthand for deriving conditional ignorability conditions.

We use the edges in a DAG to represent the inputs to the functions of a corresponding SCM. The rules for forming a DAG G_M from a SCM M are the following:

1. Represent each unobserved variable with an open vertex.
2. Represent each observed variable with a closed vertex.
3. For each assignment operation in M draw an edge from each variable on the right-hand-side of the \leftarrow operator to the variable on the left-hand-side using a solid line when both vertices are observed and a dashed line when one vertex is unobserved.

In the interest of graphical simplicity, many authors will delete nodes and edges that do not affect the results of graphical tests of interest. For example, exogenous variables that point into a single endogenous variable can often be removed.⁹ However, the removal of exogenous variables

⁹In some treatments of this material, exogenous variables always point into a single endogenous variable and dependencies among the variables are represented by dashed arcs (Pearl, 2000, Ch. 3).

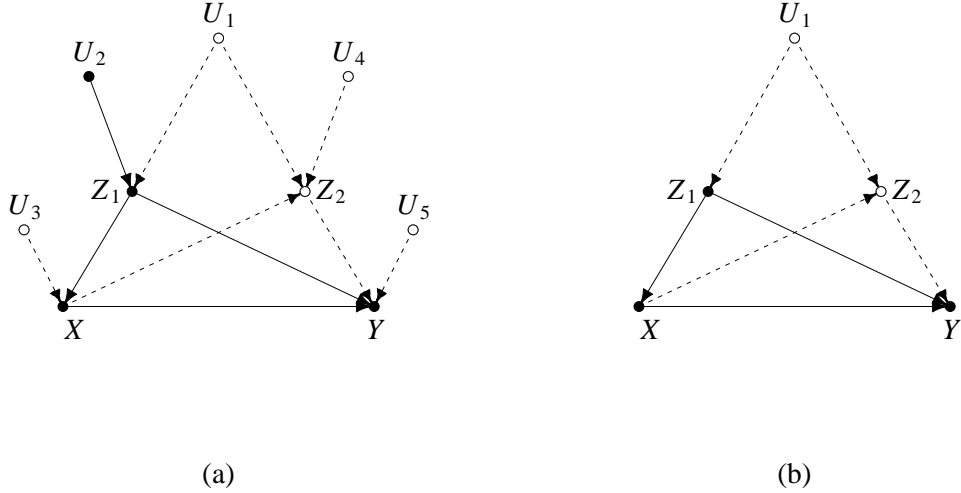


Figure 1: *Graphical Model Consistent with Structural Equations 9 - 12*. Panel (a) shows all exogenous variables and their associated edges. Panel (b) removes superfluous exogenous variables and their edges. Note that observability is neither necessary nor sufficient for a variable to be exogenous. Here U_1, \dots, U_5 are the exogenous variables. U_2 is observed but the other U variables are not. Further, one of the endogenous variables (Z_2) is unobserved while the other endogenous variables are observed.

from the graph may obscure the fact that the SCM defines individual causal effects. To avoid such confusion, in our presentation all endogenous variables have at least one exogenous variable pointing into them (with the exception of the illustrative example in Figure 1(b)).

A simple example will make this process more clear. Consider again the following structural model M :

$$z_1 \leftarrow h_{Z_1}(u_1, u_2) \tag{9}$$

$$x \leftarrow h_X(z_1, u_3) \tag{10}$$

$$z_2 \leftarrow h_{Z_2}(x, u_1, u_4) \tag{11}$$

$$y \leftarrow h_Y(x, z_1, z_2, u_5) \tag{12}$$

In Figure 1 (a), G_M is constructed using rules 1-3. A pruned version of G_M is drawn in Figure 1 (b) in which vertices and edges that are unnecessary for identifying the effect of X on Y have been dropped. The exact interpretation of the graphs in Figure 1 will wait until we discuss d -separation

in the next section.

5.3 d -Separation and Conditional Independence

Given an SCM model M and an associated causal DAG, we can read conditional independence relations from such a model with the concept of d -separation (Geiger et al., 1990).

Definition 10 (d -Separation) Let $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$ be a DAG and X, Y , and Z be disjoint subsets of \mathcal{V} . X is said to be d -separated from Y by Z in \mathcal{G} if and only if Z blocks every path from a vertex in X to a vertex in Y .

A path p is said to be blocked by a set of vertices Z if and only if at least one of the following conditions hold:

1. p contains a chain structure $a \rightarrow b \rightarrow c$ or a fork structure $a \leftarrow b \rightarrow c$ where the node b is in Z
2. p contains a collider structure $a \rightarrow b \leftarrow c$ where b is not in Z and no descendent of b is in Z

If X is not d -separated from Y by Z we say that X is d -connected to Y by Z .

The d -separation criterion is incredibly powerful in the SCM framework, because of the following theorem (Geiger et al., 1990).

Theorem 1 (Probabilistic Implications of d -Separation) Let $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$ be a DAG and X, Y , and Z be disjoint subsets of \mathcal{V} .

If Z d -separates X from Y in \mathcal{G} , then X is conditionally independent of Y given Z in every distribution compatible with \mathcal{G} .

For an SCM M , the joint distribution of the exogenous and endogenous variables is compatible with a graph G_M that is drawn using the rules from the previous subsection. Therefore, we can read conditional independence¹⁰ relations (and hence ignorability conditions) from the graph, and these form the basis of the causal identification criterion of the next subsection.

5.4 The Back-Door Criterion for Adjustment Sets

As noted in the previous subsection, there is a simple graphical criterion (Pearl, 2000) that can be checked to see if a given set \mathbf{Z} is sufficient to control confounding bias. This criterion can be stated as follows.

¹⁰Careful readers will note that the implication in Theorem 1 does not go both ways. In particular, there may be conditional independence relations in a joint distribution that are not represented by d -separation in the associated graph. However, these situations usually depend on the rare circumstances of zero effects, exact cancellation of effects, or endogenous variables that are a deterministic function of only endogenous variables. Therefore, this important caveat will not affect the results of this paper.

Definition 11 (Back-Door Criterion) *Given a causal model M and associated causal graph G_M , A set of covariates \mathbf{Z} satisfies the back-door criterion for a causal variable X and outcome Y if:*

1. \mathbf{Z} does not block any directed paths from X to (or through) Y
2. \mathbf{Z} blocks all paths from X to Y that are not directed paths

where “blocking” is defined as in Definition 10 (*d*-Separation).

If \mathbf{Z} satisfies the back-door criterion then an ignorability condition ($Y(x) \perp\!\!\!\perp X | \mathbf{Z}$) holds (Pearl, 2000), and the potential outcome distribution can be calculated using the standard stratification adjustment (Cochran, 1968; Rubin, 1977):

$$f_{Y(x)}(y) = \int_{\mathbf{z}} f_{Y|X,\mathbf{Z}}(y|x, \mathbf{z}) f_{\mathbf{Z}}(\mathbf{z}) d\mathbf{z}$$

or

$$f_{Y(x)}(y) = \sum_{\mathbf{z}} f_{Y|X,\mathbf{Z}}(y|x, \mathbf{z}) f_{\mathbf{Z}}(\mathbf{z})$$

depending on whether \mathbf{Z} is continuous or discrete and where \mathbf{Z} may be multivariate. Pearl refers to this as the *back-door* adjustment.¹¹ Since if \mathbf{Z} satisfies the back-door criterion the standard stratification adjustment is appropriate, it follows that matching or stratifying on $\Pr(x|\mathbf{z})$ (the propensity score given a realized value \mathbf{z} of \mathbf{Z}), along with related adjustments that make use of conditional ignorability, will also be appropriate (Rosenbaum and Rubin, 1983, 1984). As we will see below, this is true regardless of whether all (or even any) of the variables that affect treatment assignment are in \mathbf{Z} —all that is required is that conditional ignorability hold given \mathbf{Z} .¹²

Again, the major advantage of this graphical approach to the identification of causal effects is that it is framed in terms of a series of local assumptions about causal mechanisms. These local assumptions are often easier to consider, debate, and possibly reject as unbelievable than the single

¹¹Note that $f_{Y|X,\mathbf{Z}}(y|x, \mathbf{z})$ must exist—and thus $f_{X,\mathbf{Z}}(x, \mathbf{z})$ must be non-zero for all x and \mathbf{z} —in order for the back-door adjustment to be valid. Put slightly differently, this method of adjustment requires the distributions of the measured confounders to have the same support in the treated and control groups if an average causal effect is to be estimated. This is something that is well understood by political scientists who employ matching estimators of causal effects (Ho et al. (2007), see also King and Zeng (2006)).

¹²We note in passing the obvious point that the results of Rosenbaum and Rubin (1983) show that if conditional ignorability holds given \mathbf{Z} then using $\Pr(x|\mathbf{z})$ or any other balancing score as an adjustment covariate is appropriate. They do not show that subclassifying or matching on $\Pr(x|\mathbf{z})$ or any other balancing score for arbitrary \mathbf{Z} produces conditional ignorability of treatment assignment.

global assumption of conditional ignorability. The following examples illustrate what can go wrong when researchers use either the traditional approaches to select adjustment variables. In addition, the examples show how the back-door criterion can be easily employed to determine an appropriate adjustment strategy.

6 Examples

6.1 The Effect of British Colonial History on Authoritarianism

In our example on estimating the effects of British colonial history (BCH) on authoritarianism as encoded by polity score (PS), the researchers were willing to assume that the causal assumptions behind the linear additive model held approximately for a regression of current PS on BCH and ethnic heterogeneity as encoded by a recent ethnolinguistic fractionalization index (ELF). Using the classical econometric linear model requirements (Wooldridge, 2000; Greene, 2000) made by the researchers and using time ordering to put a causal order to the variables we can draw the causal DAG in Figure 2 (a) which is consistent with the current set of assumptions (where due to the additional parametric assumptions employed by the researchers in this example, U_1 plays the role of the classical regression error term, and the fact that the error term is exogenous can be seen because BCH and ELF are d -separated from U_1).

Using the back-door criterion, we see that *no* conditioning set that utilizes only the observed variables will allow us to identify the ACE of BCH on polity score if the world works in a way consistent with the graph in Figure 2 (a). If we condition on ELF, then we block the directed path from BCH to PS that goes through ELF. This is the familiar case of post-treatment bias (King, 1991). This case also demonstrates the inaccuracy of the usual bromide about including variables in a regression when they affect the outcome and are correlated with the explanatory variable of interest.¹³ If we instead condition on nothing (i.e. the empty set), then we leave open the back-door path from BCH through the exogenous variable U_2 , through ELF, and into PS. Informally,

¹³With all of the assumptions in this example (linearity, additivity, exogeneity), the regression parameter for BCH in the regression with ELF included will identify a partial ACE (which may or may not be of interest). However, this distinction is rarely made explicit in the political science literature.

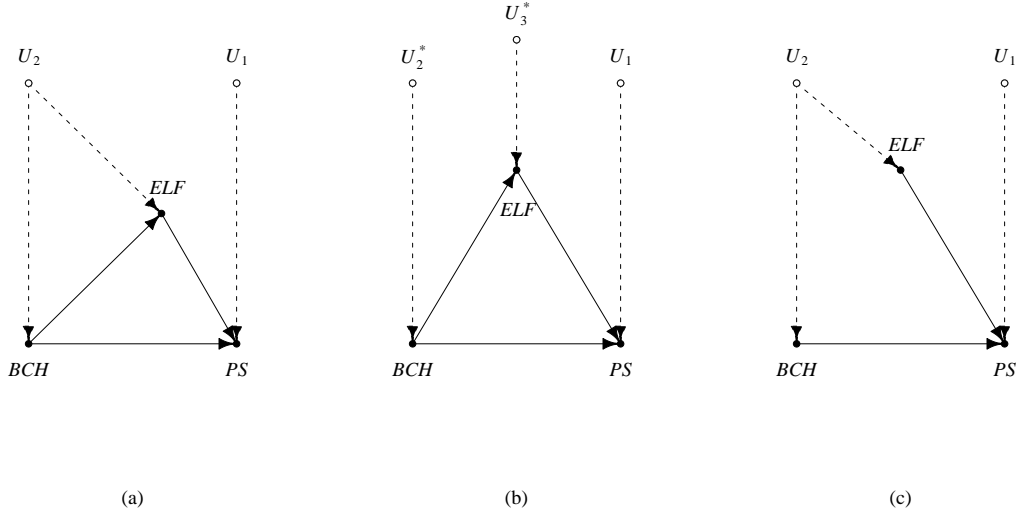


Figure 2: *Graphical Models Consistent with the Assumptions in the example on the British colonial history/authoritarianism example.* Panel (a) shows the least restrictive model (no identifying set). Panel (b) adds the assumption of independence between the background factors for BCH and ELF (the empty set is the identifying set). Panel (c) adds the assumption of no causal effect of BCH on ELF (ELF is the identifying set).

background factors are affecting both ELF and BCH, therefore omitting ELF from the regression is equivalent to standard omitted variable bias.

Given the predicament of no identifying set, the researchers might ask what further assumptions would allow identification using one of the two regressions. Figures 2 (b) and (c) show two possibilities. In Figure 2 (b), U_2 has been split into U_2^* (the background factors affecting BCH) and U_3^* (the background factors affecting ELF). Implicit in this graph is the assumption U_2^* and U_3^* are independent. If the assumptions implicit in this graph hold, then $\mathbf{Z} = \emptyset$ satisfies the back-door criterion and if the researchers' parametric assumptions also hold, the regression of PS on BCH will identify the ACE of BCH on PS. In Figure 2 (c), the edge from BCH to ELF has been removed, implying that BCH is not an input to the function that defines ELF, and hence does not affect ELF in the unit-specific effect sense. If the assumptions implicit in this graph hold, then $\mathbf{Z} = ELF$ satisfies the back-door criterion and if the researchers' parametric assumptions also hold, the regression of PS on BCH and ELF will identify the ACE of BCH on PS. Of course, neither of the additional assumptions in Figures 2 (b) and 2 (c) are likely to hold, and therefore the researchers

may want to know which regression estimator will have less bias¹⁴. Intuitively, this will depend on whether the true model is closer to Figure 2 (b) or Figure 2 (c), and a sensitivity analysis will be necessary in order to decide between the competing models.

6.2 The Effect of an Intervention on Inter-Ethnic Cooperation

Recall from our example on the effect of an intervention X on inter-ethnic cooperation as measured by an offer Y in single shot divide the dollar game, that the researchers consider two matching estimators for the ACE of X on Y , one that matches on nothing, and one that matches on the constructed survey measure of student ethnocentrism Z . Since Table 1 implies that Z is statistically dependent on X , Z conditionally dependent on Y given X , and Z is temporally prior to X , the researchers would likely judge Z to be a confounder and decide to adjust for it using exact matching or stratification (the method of adjustment is not important for this example). However, recall that while there is concern among the researchers that the ethnocentrism of a student’s parents might exert an effect on the school the child attends (and hence exposure to the program) as well as the child’s ethnocentrism and ultimately the child’s behavior in the experimental game, the researchers are also willing to assume that the background factors that influence the outcome variable (e.g. ability to recognize the strategic implications in the divide the dollar game) are independent of school choice and student ethnocentrism (and their respective background factors).

Using these assumptions and the time ordering of the variables, we can represent this model with the causal DAG in Figure 3 (a) where Z^* represents true student ethnocentrism and Z represents the reported measure of ethnocentrism. Note that the assumed independence in the background factors is represented by the separate exogenous variables (U_1 and U_2), the absence of an arrow from Z to X reflects the stated independence of U_2 and X , and the independence of U_2 and student ethnocentrism (Z^*) is maintained by the collider structure at Z . The inclusion of the remaining arrows represents the lack of additional assumptions. It is apparent from this graph and

¹⁴We use bias as the criterion here because identification and unbiasedness are equivalent in the linear model. However, for mean square error or other criteria, the necessary sensitivity analysis may be more complicated.

the back-door criterion that *no* conditioning set that utilizes only the observed variables will allow the researchers to identify the ACE of X on Y , because we do not observe Z^* , and therefore, we cannot block the back-door paths.

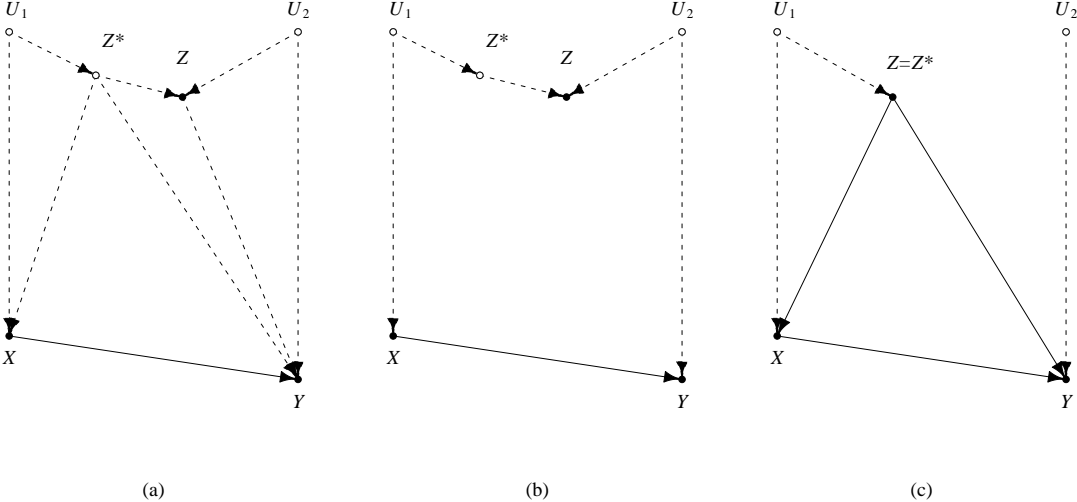


Figure 3: *Causal Graphs Consistent with the True Data Generating Process Behind Ethnic Cooperation Example (Data in Table 1)*. U_1 includes the ethnocentrism of parents, U_2 includes student IQ, Z^* is true student ethnocentrism, Z is the survey measure of student ethnocentrism, X is exposure to the program, and Y is the offer made in the experimental game. In Panel (a) there is no identifying set. In Panel (b) the empty set is the only identifying set based on observed variables. In Panel (c), $\{Z^*\}$ is the only identifying set based on observed variables.

If we make the additional assumptions implicit in Figure 3 (b), then the empty set is sufficient to block all back door paths, and conditioning on Z would lead to bias. *Therefore, conditioning on an irrelevant pre-treatment variable induces bias where there was none before.* Hence, this example directly contradicts the received wisdom from the classical econometrics literature. To see the intuition behind this, suppose that ethnicity has no effect on the offers in the game (even through the intervention at the inclusive school). In this case we could further remove the arrow from X to Y , and the unit-specific effect of X on Y is zero for all students. As the researchers expect, parental ethnocentrism (U_1) does affect the survey measure of student ethnocentrism (Z) and exposure to the program (X). What the researchers failed to recognize is that the survey measure of ethnocentrism is also affected by a second latent variable— “student IQ” (U_2)—and this latent variable also has

an effect on the outcome variable (Y). The mechanisms here are the following. Students with higher IQs are better able to grasp what the survey instruments are probing. If students realize that the instrument is attempting to measure a socially undesirable trait and they possess that trait they will misreport their true beliefs on the survey. Further, students with higher IQs are more likely to realize the monetary incentives embedded in the experimental game and offer \$0.01. Importantly, the missing edges from Z to X and from Z to Y imply the lack of the corresponding direct effects. In this scenario, knowing that a student goes to the ethnocentric school tell us that they are more likely to have ethnocentric parents, and are themselves more likely to be ethnocentric, but it tells us nothing about their “student IQ” (U_2). Therefore, ignorability holds because the potential outcomes are determined by U_2 in this simplified model, and treatment assignment tell us nothing about the potential outcomes. However, if we know that a student at the ethnocentric school reported low ethnocentricity ($Z = 0$), then we know that the low reported ethnocentricity is more likely to be a result of high “student IQ” (U_2). Therefore, knowing treatment assignment gives us information about the potential outcomes, and ignorability is violated.

Table 2 presents two estimated potential outcome distributions that are derived from the distribution in Table 1 which is consistent with the causal DAG in Figure 3 (b). The putative estimates are based on adjusting for Z . This yields the potential outcome distributions in the two leftmost columns of Table 2. From this table it appears that the program exerts a modest, positive effect on cooperative behavior. The counterfactual probability of the most selfish offer ($Y = \$0.01$) decreases by 3 percentage points under treatment while the counterfactual probability of the most equitable offer ($Y = \$10.00$) increases by the same amount. The true potential outcome distributions are shown in the two rightmost columns of Table 2. Here we see that there is no causal effect of the program on outcomes.

It is important to note that the consistency of $\Pr(y|x)$ for $\Pr(Y(x) = y)$ and the inconsistency¹⁵

¹⁵Inconsistency statements of this type require the additional assumption of the *faithfulness* (Spirtes et al., 1993) of the graph (i.e. the graph encodes all conditional independence relations in the population distribution). But in this context, such an assumption is actually conservative and will hold for all but the most contrived distributions.

	Putative $\Pr(Y(0) = y)$	Putative $\Pr(Y(1) = y)$	True $\Pr(Y(0) = y)$	True $\Pr(Y(1) = y)$
$y = \$0.01$	0.52	0.49	0.50	0.50
$y = \$10.00$	0.43	0.46	0.45	0.45
$y = \$19.99$	0.05	0.05	0.05	0.05

Table 2: *Putative Potential Outcome Distribution Based on Data in Table 1 After Adjusting for Z along with the True Potential Outcome Distribution.*

of estimators that adjust for Z does not depend on the type of adjustment method, particular parametric assumptions, or unlikely cancellations of effects. It is a general result for all causal models that are compatible with the DAG in Figure 3 (b); or even more generally for all causal models in which the back-door criterion holds for \emptyset but not for \mathbf{Z} .

If instead of the assumptions implicit in Figure 3 (b) we make the assumptions implicit in Figure 3 (c), then Z is a perfect copy of Z^* (i.e. students report their true ethnocentrism), and we can block all back door paths by conditioning on Z . However, because the assumptions implicit in Figures 3 (b) and Figure 3 (c) are not likely to hold exactly, the researchers may want to know which regression estimator will be more accurate. Intuitively, this will depend on whether the true model is closer to 2 (b) or 2 (c), and a sensitivity analysis will be necessary in order to decide between the competing models.

6.3 A More Complicated Parametric Example

As noted above, the back-door criterion is relevant for a wide range of commonly-used adjustment methods within a wide range of observational and / or experimental studies. This is true even when the underlying causal relationships are much more complicated than the stylized example in the previous subsection. In this subsection, we demonstrate these points by using the back-door criterion to identify sets of covariates sufficient to control confounding in a more complicated example. We then show how two widely used methods of adjustment—linear regression and subclassification on the estimated propensity score—can be used to estimate average causal effects as well as how these methods fall prey to the same problems discussed in the previous subsection.

Consider the graphical causal model in Figure 4. Here we see four measured background variables (Z_1, Z_2, Z_3 , and Z_4), a single treatment variable (X), an intermediate outcome variable (Z_5), and a final outcome variable (Y). In addition, there are numerous unmeasured exogenous variables (U_1, U_2, \dots, U_{11}).

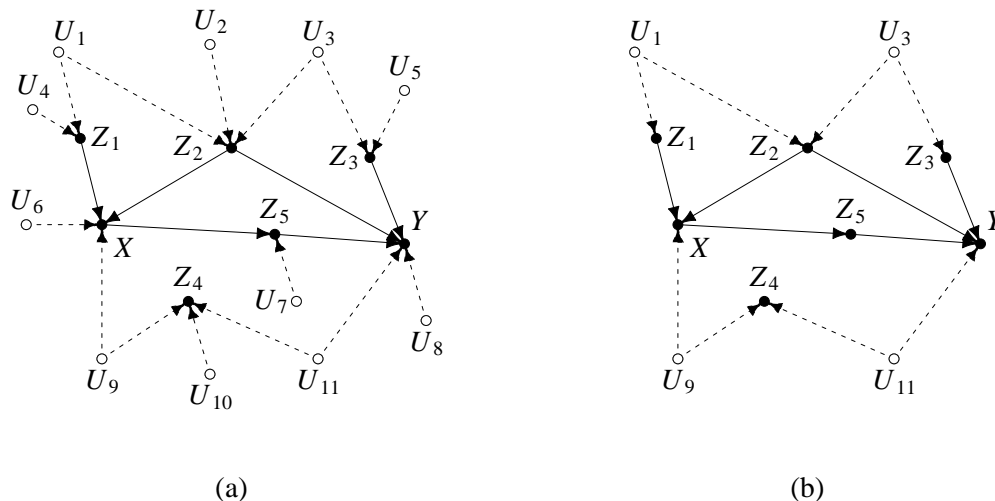


Figure 4: *Causal Graphs Consistent with Linear Structural Equations in Figure 5.* Panel (a) contains all exogenous variables while panel (b) excludes nodes and edges that are unnecessary for determining if and how the causal effect of X on Y is identified.

The research question of interest is to determine the average causal effect of X on Y . To identify the sets of observed covariates that are sufficient for the control of confounding we check the back-door criterion for all subsets of the measured covariates. Here we see that $\{Z_1, Z_2\}$, $\{Z_2, Z_3\}$, and $\{Z_1, Z_2, Z_3\}$ are the sets that satisfy the back-door criterion. Thus adjusting for $\{Z_1, Z_2\}$, $\{Z_2, Z_3\}$, or $\{Z_1, Z_2, Z_3\}$ will provide a consistent estimate of the causal effect of X on Y .¹⁶ Note two interesting things about this example. First, the sufficient sets need not be nested. Second, Z_4 is not a member of any of the sets of adjustment variables above. Thus it should never be conditioned on— *even though it is associated with both X and Y given X and could be a pre-treatment variable.*

To make this discussion more concrete, we assume particular linear forms for the structural

¹⁶The structure of the causal graph in Figure 4 also suggests a fourth way to identify the causal effect of X on Y by using the intermediate outcome Z_5 . This is the so-called *front-door adjustment*, which is outside the scope of this paper.

equations that are consistent with the causal graph in Figure 4. These linear structural equations are presented in Figure 5. In what follows, we generate 1,000,000 independent replicates from this system of equations and then use linear regression and subclassification on the estimated propensity score to estimate the causal effect of X on Y . More specifically, we show how in each case adjustment for $\{Z_1, Z_2\}$, $\{Z_2, Z_3\}$, or $\{Z_1, Z_2, Z_3\}$ provides an accurate estimate of the causal effect of X on Y , and how in each case conditioning on an incorrect set of adjustment variables produces a very inaccurate estimate of the effect of interest. Before proceeding, we note that the linearity of the equations in Figure 5 allows us to easily derive the true average causal effect $\mathbb{E}[Y(X = 1)] - \mathbb{E}[Y(X = 0)]$ from the parameter values in these equations. Here we see that the true average causal effect is 0.5.

$$\begin{array}{ll}
u_1 \stackrel{ind}{\sim} \mathcal{N}(0, 0.75^2) & u_7 \stackrel{ind}{\sim} \mathcal{N}(0.6, 1^2) \\
u_2 \stackrel{ind}{\sim} \mathcal{N}(0, 3^2) & u_8 \stackrel{ind}{\sim} \mathcal{N}(1, 1^2) \\
u_3 \stackrel{ind}{\sim} \mathcal{N}(5, 0.4^2) & u_9 \stackrel{ind}{\sim} \mathcal{N}(0, 5^2) \\
u_4 \stackrel{ind}{\sim} \mathcal{N}(0, 1^2) & u_{10} \stackrel{ind}{\sim} \mathcal{N}(0, 0.05^2) \\
u_5 \stackrel{ind}{\sim} \mathcal{N}(1, 0.8^2) & u_{11} \stackrel{ind}{\sim} \text{Binom}(4, 0.55) \\
u_6 \stackrel{ind}{\sim} \mathcal{N}(0, 5^2) & \\
\\
z_1 \leftarrow u_1 + u_4 & z_3 \leftarrow u_3 + u_5 \\
z_2 \leftarrow 0.5u_1 + u_2 + u_3 & z_4 \leftarrow -0.2u_9 + u_{10} - 0.25u_{11} \\
\\
x \leftarrow \begin{cases} 1 & \text{if } z_1 + 2.5z_2 + u_6 - u_9 > 12.5 \\ 0 & \text{otherwise} \end{cases} \\
\\
z_5 \leftarrow x + u_7 \\
y \leftarrow 0.5z_5 + z_2 - 1.5z_3 + u_8 - 2u_{11}
\end{array}$$

Figure 5: *Specific Linear Structural Equations Consistent with Figure 4.* The results discussed in this section are based on 1,000,000 observations generated from this model.

Consider the estimation of $\mathbb{E}[Y(X = 1)] - \mathbb{E}[Y(X = 0)]$ using linear regression. While linear regression will typically not produce a consistent estimate of the potential outcome distribution,

it will produce a consistent estimate of the average causal effect if the system is linear and the back-door criterion is satisfied for some subset of measured covariates (Pearl, 2000, Chapter 5). Thus we would expect that the estimates of β_1 in the following regression equations

$$y = \beta_0 + \beta_1 x + \beta_2 z_1 + \beta_3 z_2 + \epsilon$$

$$y = \beta_0 + \beta_1 x + \beta_2 z_2 + \beta_3 z_3 + \epsilon$$

$$y = \beta_0 + \beta_1 x + \beta_2 z_1 + \beta_3 z_2 + \beta_4 z_3 + \epsilon$$

would all be close to the true value of 0.5. Looking at Table 3 we see that this is in fact the case. We also see that, as we would expect, omitting Z_2 from $\{Z_1, Z_2, Z_3\}$ and only conditioning on Z_1 and Z_3 results in an estimated causal effect that is far from the truth.

Finally, note that when all of the observed covariates are included as right-hand-side variables in a regression model that the sign of the estimated causal effect is the opposite of that of the truth. While the sign and magnitude of this bias depend on the specific forms of the structural equations in Figure 5, the fact that conditioning on all the observed covariates ($\{Z_1, Z_2, Z_3, Z_4\}$) results in an inconsistent estimate of the average causal effect only depends on the general relationships embodied in the graphical structure in Figure 4.

Some intuition about what is happening in this linear example is the following. First note that the disturbance in the equation for Y always includes U_{11} . Marginally, X and U_{11} are independent. However, *given* Z_4 , X and U_{11} are generically *dependent*. More specifically, given the equations for X and Z_4 in Figure 5 we see that Z_4 will likely take a relatively large positive value when either U_9 , U_{11} , or both take a large negative value. Further, the probability of $X = 1$ is decreasing in realized values of U_9 . Thus, conditioning on a large positive value of Z_4 and observing that $X = 0$ provides evidence that U_{11} has taken a large negative value. Thus, when Z_4 is included as a right-hand-side variable in a regression model of Y on X the disturbance term (which is always defined conditionally on all covariates in the model) is no longer independent of X . As a result, the OLS estimator of the coefficient on X is no longer consistent for the total causal effect.

Adjustment Model	$\hat{\mathbb{E}}[Y do(X=1)] - \hat{\mathbb{E}}[Y do(X=0)]$
$y = \beta_0 + \beta_1x + \beta_2z_1 + \beta_3z_2 + \epsilon$	0.504
$y = \beta_0 + \beta_1x + \beta_2z_2 + \beta_3z_3 + \epsilon$	0.505
$y = \beta_0 + \beta_1x + \beta_2z_1 + \beta_3z_2 + \beta_4z_3 + \epsilon$	0.505
$y = \beta_0 + \beta_1x + \beta_2z_1 + \beta_3z_3 + \epsilon$	4.063
$y = \beta_0 + \beta_1x + \beta_2z_1 + \beta_3z_2 + \beta_4z_3 + \beta_5z_4 + \epsilon$	-0.179
$\Pr(X=1 z) = \Phi(\alpha_0 + \alpha_1z_1 + \alpha_2z_2)$	0.521
$\Pr(X=1 z) = \Phi(\alpha_0 + \alpha_1z_2 + \alpha_2z_3)$	0.531
$\Pr(X=1 z) = \Phi(\alpha_0 + \alpha_1z_1 + \alpha_2z_2 + \alpha_3z_3)$	0.522
$\Pr(X=1 z) = \Phi(\alpha_0 + \alpha_1z_1 + \alpha_2z_3)$	4.063
$\Pr(X=1 z) = \Phi(\alpha_0 + \alpha_1z_1 + \alpha_2z_2 + \alpha_3z_3 + \alpha_4z_4)$	-0.218
Truth	0.500

Table 3: *Adjustment Methods and Associated Estimated Average Causal Effects for Example Consistent with Figures 5 and 4.* The first five rows correspond to various regression adjustments. The second five rows correspond to various probit regressions for the estimated propensity scores. In these rows, the estimated causal effect is calculated by subclassifying on the estimated propensity score. The true average causal effect is in the last row.

The same general patterns of bias emerge when one estimates the average causal effect of X on Y by subclassifying on the estimated propensity score. Since the entire system of equations is linear (and the true data generating process for X is consistent with a probit model), we use generic probit regression with main effects for the conditioning variables to estimate the propensity score. The various specifications for the propensity score model are given in the second five rows of Table 3. In general, the patterns of bias are identical to those seen in the equivalent linear model adjustments.¹⁷ This should be of no surprise, since when the data are generated according to a linear model both methods are doing very similar things. Nevertheless, there are some interesting points to note here. First, as most researchers realize (and is made clear in the seminal work of Rosenbaum and Rubin (1983)) the propensity score model need not be consistent with the true data generating process for the treatment variable or even include that true data generating process as a special case of the estimated propensity score model. It is easy to prove that all that is required is that conditional ignorability holds given \mathbf{z} and that the statistical model used to estimate $\Pr(x|\mathbf{z})$ is sufficiently flexible to accurately estimate this distribution. The fact that subclassifying on the

¹⁷The slight upward bias of most of the propensity score estimates is likely do to the fact that these estimates were based on a less than optimal subclassification plan.

estimated propensity scores from the propensity score model:

$$\Pr(X = 1|z) = \Phi(\alpha_0 + \alpha_1 z_2 + \alpha_2 z_3)$$

provide essentially the same estimate as subclassifying on the propensity scores from the probit model that is consistent with the data generating process confirms this point.

What will be more surprising to many readers is the fact that subclassifying on the propensity scores from the model:

$$\Pr(X = 1|z) = \Phi(\alpha_0 + \alpha_1 z_1 + \alpha_2 z_2 + \alpha_3 z_3 + \alpha_4 z_4)$$

produces a badly biased estimate with the wrong sign. This propensity score model includes the true data generating model for X as a special case (with α_3 and α_4 both equal to 0) as well as all the other propensity score models that would allow one to construct consistent estimates of the causal effect of interest. Further, all of the covariates can be thought of as pre-treatment variables with no loss of generality. Thus many researchers might conclude there is no real harm to conditioning on all the observed covariates. This is not correct. Conditioning on all the observed covariates produces a misleadingly biased estimate of the causal effect in this example.

7 Conclusion

In this paper we have attempted to make the case that the SCMs of Pearl (a) are consistent with how many empirical political scientists think about causality within their subject-matter area, (b) are equivalent to the powerful and well-respected Neyman-Rubin Causal Model, and (c) offer the potential to improve the practice of causal inference in political science. Political scientists tend to think in terms of causal mechanisms and the SCM deals explicitly in mechanisms. The SCM is a model of deterministic potential outcomes and the key ideas and results from the Neyman-Rubin model have direct analogies in this framework. The SCM offers several advantages to applied researchers. We discuss each of these below.

The SCM provides explicit rules for covariate selection based on the causal graph. Given a graph (or a set of plausible graphs), the back-door criterion supplies a collection of sufficient conditioning sets to identify the ACE, thus eliminating (and explaining) the inaccuracies in the traditional regression criteria. Furthermore, the transparent local assumptions about mechanisms in this model are easier to consider, debate, and potentially reject than a global assumption of conditional ignorability. Causal inference by its very nature relies on untestable causal assumptions. As such it is imperative that any method for causal inference allow researchers to state these causal assumptions as plainly and comprehensibly as possible so that subject-matter experts can easily weigh in on the plausibility of these assumptions. Finally, SCMs provide structure that can inform the design of future studies and sensitivity analyses when no identifying adjustment set can be found. We will address these issues in future work.

While the SCM provides the advantages described above, we emphasize that it is consistent with and can be used in conjunction with the the Neyman-Rubin framework that has been so useful to political scientists. Furthermore, while the assumptions sufficient for the identification of ACEs are easier to assess with causal DAGs, the assumptions sufficient for the identification of other causal effects may be easier to assess with the pure Neyman-Rubin framework (see the discussion of Pearl (1995)). In many situations a hybrid approach that utilizes the strengths of both models seems appropriate.

References

- Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91:444–455.
- Campbell, Angus, Philip Converse, Warren E. Miller, and Donald Stokes. 1960. *The American Voter*. New York: John Wiley.
- Clarke, Kevin A. 2005. "The Phantom Menace: Omitted Variable Bias in Econometric Research." *Conflict Management and Peace Science* 22:341–352.
- Clarke, Kevin A. 2006. "Return of the Phantom Menace: Omitted Variable Bias in Political Research." .
- Cochran, William G. 1968. "The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies." *Biometrics* 24(2):295–313.
- Collier, David, and Henry E. Brady. 2004. *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. Lanham, MD: Rowman & Littlefield.
- Cox, Gary W. 1997. *Making Votes Count: Strategic Coordination in the World's Electoral Systems*. Cambridge: Cambridge University Press.
- Elster, Jon. 1989a. *The Cement of Society*. Cambridge: Cambridge University Press.
- Elster, Jon. 1989b. *Nuts and Bolts for the Social Sciences*. Cambridge: Cambridge University Press.
- Elster, Jon. 1998. "A Plea for Mechanisms." In *Social Mechanisms: An Analytical Approach to Social Theory* (Peter Hedström, and Richard Swedberg, editors), Cambridge: Cambridge University Press.
- Fearon, James D. 1995. "Rationalist Explanations for War." *International Organization* 49(3):379–414.
- Fox, John. 1997. *Applied Regression Analysis, Linear Models, and Related Methods*. Thousand Oaks, CA: Sage.
- Geiger, D., T. Verma, and J. Pearl. 1990. "Identifying independence in Bayesian networks." *NETWORKS*. 20(5):507–534.
- Greene, William H. 2000. *Econometric Analysis*. New York: Macmillan, fourth edition.
- Ho, Daniel, Kosuke Imai, Gary King, and Elizabeth Stuart. 2007. "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." *Political Analysis* 15:199–236.
- Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81:945–960.
- King, G. 1991. "'Truth' Is Stranger than Prediction, More Questionable than Causal Inference." *American Journal of Political Science* 35(4):1047–1053.
- King, Gary, and Langche Zeng. 2006. "The Dangers of Extreme Counterfactuals." *Political Analysis* 14(2):131–159.
- Kmenta, Jan. 1986. *Elements of Econometrics*. New York: Macmillan, second edition.

- Neyman, Jerzy, (with K. Iwazskiewicz, and S. Kolodziejczyk). 1935. "Statistical Problems in Agricultural Experimentation." *Supplement of Journal of the Royal Statistical Society* 2:107–180.
- Pearl, Judea. 1995. "Causal Diagrams for Empirical Research." *Biometrika* 82:669–710.
- Pearl, Judea. 2000. *Causality: Models, Reasoning, and Inference*. New York: Cambridge University Press.
- Pindyck, Robert S., and Daniel L. Rubinfeld. 1998. *Economic Models and Economic Forecasts*. Boston: Irwin McGraw Hill, fourth edition.
- Robins, J.M. 1986. "A new approach to causal inference in mortality studies with a sustained exposure period-application to control of the healthy worker survivor effect." *Mathematical Modeling* 7:1393–1512.
- Rosenbaum, Paul R., and Donald B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70:41–55.
- Rosenbaum, Paul R., and Donald B. Rubin. 1984. "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score." *Journal of the American Statistical Association* 79:516–524.
- Rothman, Kenneth J. 1986. *Modern Epidemiology*. Boston: Little Brown.
- Rothman, Kenneth J., and Sander Greenland. 1998. *Modern Epidemiology*. Philadelphia: Lippincott-Raven.
- Rubin, D.B. 2004. "Direct and Indirect Causal Effects via Potential Outcomes*." *Scandinavian Journal of Statistics* 31(2):161–170.
- Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66(5):688–701.
- Rubin, Donald B. 1977. "Assignment to Treatment Group on the Basis of a Covariate." *Journal of Educational Statistics* 2(1):1–26.
- Rubin, Donald B. 1978. "Bayesian Inference for Causal Effects: The Role of Randomization." *The Annals of Statistics* 6(1):34–58.
- Schlesselman, James J. 1982. *Case-Control Studies: Design Conduct Analysis*. New York: Oxford University Press.
- Spirtes, P., C. Glymour, and R. Scheines. 1993. *Causation, Prediction, and Search*. New York: Springer.
- Wooldridge, J.M. 2000. "Introductory Econometrics: A Modern Approach, 2e." *Thomson South-Western, USA* .