Advance Access publication July 19, 2011

Political Analysis (2011) 19:273–286
doi:10.1093/pan/mpr021

Why Process Matters for Causal Inference

Adam N. Glynn
Department of Government, Harvard University, 1737 Cambridge Street, Cambridge, MA 02138
e-mail: aglynn@fas.harvard.edu (corresponding author)

Kevin M. Quinn
UC Berkeley School of Law, 490 Simon 7200, Berkeley, CA 94720-7200
e-mail: kquinn@law.berkeley.edu

Our goal in this paper is to provide a formal explanation for how within-unit causal process information (i.e., data on posttreatment variables and partial information on posttreatment counterfactuals) can help to inform causal inferences relating to total effects—the overall effect of an explanatory variable on an outcome variable. The basic idea is that, in many applications, researchers may be able to make more plausible causal assumptions conditional on the value of a posttreatment variable than they would be able to do unconditionally. As data become available on a posttreatment variable, these conditional causal assumptions become active and information about the effect of interest is gained. This approach is most beneficial in situations where it is implausible to assume that treatment assignment is conditionally ignorable. We illustrate the approach with an example of estimating the effect of election day registration on turnout.

1 Introduction

Simple versions of the so-called "Neyman–Rubin causal model" take an essentially black-box approach to causality—in an ideal randomized experiment, it is not necessary to know how, that is, through what mechanisms, a particular treatment works in order to consistently estimate a variety of causal effects. All that is necessary is random assignment of treatment along with assumptions that guarantee that potential outcomes are well defined for all relevant units. Of course, scholars as diverse as Fisher (quoted in Cochran 1965; Collier and Brady 2004; George and Bennett 2005; Brady, Collier, and Seawright 2006; Hedström 2008; Deaton 2009; Heckman and Urzua 2009, and others) have argued that one can, and should, posit causal mechanisms linking the treatment variable to the outcome variable and one should think carefully about the observable implications of this casual process. This is especially important when confronted with observational data in which the treatment was not randomized.

Our goal in this paper is to provide a formal explanation for how within-unit causal process information (i.e., data on posttreatment variables and partial information on posttreatment counterfactuals) can help to inform causal inferences relating to total effects—the overall effect of an explanatory variable on an outcome variable. The basic idea is that, in many applications, the treatment has not been randomized, and therefore researchers will have to make assumptions in order to identify causal effects (even on average). As we demonstrate in this paper, researchers may be able to make more plausible causal assumptions conditional on the value of a posttreatment variable than they would be able to do unconditionally (or conditionally on the measured pretreatment variables). As data become available on a posttreatment variable, these conditional causal assumptions become active and information about the total effect is gained.

Authors' note: The authors thank Matt Chingos for his research assistance and Kevin Clarke, Luke Keele, Jay Kaufman, James Mahoney, the participants of the 2009 meeting of the Midwest Political Science Association, the participants of the 2009 Causal Workshop at the Banff International Research Station, the participants of the 2009 Summer meeting of the Society of Political Methodology, two anonymous referees, and the editors for their helpful comments and suggestions.

Note that this formal explanation is necessary to reconcile the aforementioned advice (that one should utilize posttreatment variables in an analysis of total effects) with the standard warnings about posttreatment bias (which is induced by adjusting for a posttreatment variable in a regression or matching analysis; Rosenbaum and Rubin 1983; King 1991; King, Keohane, and Verba 1994; Pearl 2009, Section 3.3.1). Both pieces of advice are correct, and the formal explanation presented in this paper demonstrates how to utilize posttreatment variables without inducing posttreatment bias.

To get a sense of the argument, consider the following stylized example. We are interested in the effect of election day registration (EDR) laws on turnout. We observe data from a single individual, and we know that in 2004, she did not live in a state with EDR available and that she did not vote in the 2004 presidential election. Without additional knowledge, all that we can say is that the individual causal effect of EDR for this person is either 0 (she is a "never voter" and would not have voted even if EDR had been available) or it is positive (she would have been "helped" to vote by the availability of EDR). Now, suppose we learn that this person registered to vote thirty days prior to the election in 2004. Since most, if not all, of the presumed effect of EDR provisions on voting is assumed to work by decreasing the costs of registration and thus increasing registration, it seems that most researchers would now be more likely to conclude that our hypothetical citizen is a "never voter"—in other words, enacting an EDR provision in this person's state in 2004 would not have increased her likelihood of voting since she was already registered on election day. Obviously, this conclusion depends on strong assumptions (although these assumptions seem plausible for this scenario), and for many applications, we may want to weaken these assumptions (or combine them with other types of assumptions). Furthermore, although this story is operating at the individual level, it is not difficult to aggregate over a number of such individual-level stories in order to estimate the distribution of effects and/or average effects.

A small but growing literature attempts to formalize the use of posttreatment/process information to improve inference for total causal effects. Rosenbaum (1984) demonstrated that conditioning on posttreatment variable could sometimes reduce bias when that variable was a surrogate for a pretreatment variable. Pearl (1995, 2009) and its extensions (Tian and Pearl 2002a, 2002b) went further, demonstrating that when the front-door or single-door criteria hold, the average total causal effect could be nonparametrically identified with the use of posttreatment variables, even when treatment assignment is nonignorable. More recent work has focused on cases when the assumptions necessary for the front-door technique do not fully hold. VanderWeele (2008) and VanderWeele and Robins (2010) demonstrated that posttreatment variables might be used to determine the direction of bias. Joffe (2001) finds bounds for total effects that utilize the observation of a posttreatment variable, and Kaufman, Kaufman, and MacLehose (2009) uses linear programming via the OPTIMIZE program (Balke 1995) to provide a number of bounds based on alternative sets of assumptions (some of which involve the observation of a posttreatment variable).[1]

This paper contributes to the "bounds" strand of this literature, demonstrating (with an example from political science) how knowledge of the *empirical* joint distribution of a treatment variable, a posttreatment variable, and the outcome variable, combined with plausible assumptions regarding causal effects within some, but not necessarily all, strata of this joint distribution can narrow the usual Manski (2003) bounds for the effect of interest. This approach is most beneficial in situations where it is implausible to assume that treatment assignment is conditionally ignorable. We illustrate the approach with an example of estimating the effect of EDR on turnout. The utility of our approach is highlighted by the fact that once one thinks seriously about what knowledge of the posttreatment variable (here registration) implies, it becomes apparent that the effect estimates produced by standard data analytic strategies are implausibly high.

[1]There is also a great deal of related work on different causal estimands within the same general class of models (often when the treatment is randomly assigned or randomly encouraged). For example, instrumental variables estimators (Angrist, Imbens, and Rubin 1996) and principal stratification (Frangakis and Rubin 2002) utilize the same model, and recent work has related the intent-to-treat effects conditional on principal strata to as-treated effects conditional on the observed compliance behavior (TenHave et al. 2004). Furthermore, work on indirect effects and mediation analysis in the potential outcomes framework uses this same model (Robins and Greenland 1992; Pearl 2001; Robins 2003; Glynn 2009; Imai et al. 2010 Pearl 2010).

2 A Framework for Reasoning about Causal Process Information

2.1 *Causal Inference with Randomized Treatment Assignment: A Review*

Much of causal inference centers on questions regarding treatment effects (sometimes known as total effects). At the unit level, we would write such effects for units $i = 1, \ldots, n$ as the difference between the observed outcome $Y_i \equiv Y_i(X_i)$ and a counterfactual outcome (as long as the counterfactual is assumed to be well defined). For example, if $X_i = x$, then the unit-level treatment effect of changing from x to x' can be written as the following:

$$Y_i(x') - Y_i(x).$$

The two values in this contrast are often called potential outcomes, and this model for causal effects is sometimes known as the potential outcomes model or the Neyman–Rubin causal model (Neyman 1923; Copas 1973; Rubin 1978).

Unfortunately, although the potential outcome $Y_i(x)$ is observable when $X_i = x$, the other potential outcome, $Y_i(x')$, is not observable when $x \neq x'$ because $Y_i(x')$ is counterfactual. Analogously, when $X_i = x'$, $Y_i(x')$ is observable but $Y_i(x)$ is not. The fact $Y_i(x)$ and $Y_i(x')$ can never be simultaneously observed for the same unit is sometimes known as the fundamental problem of causal inference (FPOC; Holland 1986).

It is now well known that the most reliable approach to partially solving the FPOC involves randomly assigning units to the values x and x' (although there are many different methods of randomization and some may be more reliable than others depending on the application). In its most simple form, randomization of units to x and x' allows random draws from the marginal distributions of the potential outcomes within the sample of $i = 1, \ldots, n$ units. If the n sample units are themselves randomly drawn from the population, then randomization of units to x and x' allows random draws from the marginal population distributions, which we denote as $\mathbb{F}_{Y(x)}$ and $\mathbb{F}_{Y(x')}$. Thus, as $n \to \infty$, we get consistent estimates of the marginal population distributions of the potential outcomes.

The FPOC implies that randomization does not provide draws from the joint distribution of potential outcomes ($\mathbb{F}_{Y(x),Y(x')}$) or the distribution of effects ($\mathbb{F}_{Y(x')-Y(x)}$), however, randomization provides point identification for average effects such as the average treatment effect (ATE) (also known as the average total effect):

$$\begin{aligned} \text{ATE} &\equiv \mathbb{E}[Y(x') - Y(x)] \\ &= \mathbb{E}[Y(x')] - \mathbb{E}[Y(x)] \end{aligned}$$

and conditional average treatment effects (CATE) (also known as the conditional average total effects),

$$\begin{aligned} \text{CATE} &\equiv \mathbb{E}[Y(x') - Y(x)|W = w] \\ &= \mathbb{E}[Y(x')|W = w] - \mathbb{E}[Y(x)|W = w]. \end{aligned}$$

It is important to note that W should not be posttreatment, but for many applications, W may be the treatment itself $W = X$. An example of this will be presented in Section 3.

2.2 *Assumptions Based on Pretreatment Variables: A Review*

In the absence of randomization, one must use additional assumptions for the identification of ATE or CATE. For example, suppose we want to know the average effect for the units $i = 1, \ldots, n_x$ for whom $X_i = x$. For these units, $Y_i(x)$ is observed as Y_i, whereas $Y_i(x')$ is unobserved, so for the average effect of interest,

$$\begin{aligned} \mathbb{E}[Y(x') - Y(x)|X = x] &= \mathbb{E}[Y(x')|X = x] - \mathbb{E}[Y(x)|X = x] \\ &= \mathbb{E}[Y(x')|X = x] - \mathbb{E}[Y|X = x] \end{aligned}$$

the quantity $\mathbb{E}[Y(x')|X = x]$ can only be estimated by making assumptions. Nearly always, these assumptions require the measurement of pretreatment variables in order to make them operational.

For example, one very popular assumption is that conditional on a pretreatment variable (or set of variables) W, the assignment of treatment is ignorable (Rosenbaum and Rubin 1983). Intuitively, this means that the units with $X_i = x$ and $W_i = w$ can be compared directly to units with $X_i = x'$ and $W_i = w$, so that $\mathbb{E}[Y(x')|X = x]$ can be estimated by averaging the Y_i values for units with $X_i = x'$ and $W_i = w$. The key point is that without measuring W, it is not possible to utilize this assumption for inference. In other words, the measurement of W makes the ignorability assumption operational.

Another typical example uses a pretreatment variable known as an instrument and involves the use of alternative assumptions (Angrist et al. 1996). In the simplest case, a binary instrument (Z) is assumed to have been assigned ignorably, the effect of Z on a binary X is assumed to be monotonic (and to exist for at least some units), and the effect of Z on a binary outcome (Y) is assumed to be due entirely to the effect of Z on X (i.e., an exclusion restriction holds).[2] If these assumptions hold, then although $\mathbb{E}[Y(x')|X = x]$ cannot be estimated precisely (without additional assumptions), the bounds of possible estimates may be substantively informative (Pearl 2009). Again, the key point is that without measuring Z, it is not possible to utilize these assumptions for inference. In this sense, measurement of Z provides information about the causal effect when the instrumental variables assumptions are plausible.

2.3 *Assumptions Based on Posttreatment Variables*

The remainder of this paper will demonstrate that posttreatment variables can provide information about causal effects in a manner similar to pretreatment variables. To demonstrate this formally, consider a posttreatment variable M (we could alternatively consider M to be a set of posttreatment variables, but for simplicity in presentation, we will assume M to be a scalar). Sometimes posttreatment variables are known as process variables, intermediate variables, or mediating variables. Often the processes described by these variables are known as mechanisms.

As with the potential outcomes defined above, we can define the effect of X on M in terms of potential mediators. For example, if $X_i = x$, then $M_i = M_i(x)$ and if the counterfactual mediator $M_i(x')$ is well defined, then the unit-level treatment effect of changing from x to x' can be written as the following:

$$M_i(x') - M_i(x),$$

where $M_i(x) = M_i$ is observed and $M_i(x')$ is unobserved. If the potential mediator $M_i(x)$ is well defined, and M is part of the process or mechanism by which X affects Y, then the potential outcome $Y(x)$ can be written as the following:

$$Y_i(x) \equiv Y(x, M_i(x)),$$

which can be interpreted as the value of Y that we would observe if X had been x and M had been what it would have been if X had been x. If the potential mediator $M_i(x')$ is well defined, the potential outcome $Y_i(x')$ can be defined analogously as $Y(x', M_i(x'))$.

It is important to note what we have not assumed. We have not assumed that quantities such as $Y(x, M_i(x'))$ or $Y(x', M_i(x))$ (which involve the values x and x' simultaneously) are well defined. These types of quantities are necessary for path analysis (also known as mediation analysis)—the attempt to decompose total causal effects into path specific effects—but they are not necessary when using process analysis to learn about total effects (the goal of this paper). Furthermore, we have not assumed that $Y(x, M_i(x)) = Y(x, m)$ when $M_i(x) = m$. In words, we do not need to assume that the potential outcome would be the same regardless of how we set the value of the mediator. For this paper, we need only consider cases when the mediating variable is set by the process of intervening on X.

To see why the consideration of posttreatment variables can make additional assumptions operational, consider again making inferences about the quantity $\mathbb{E}[Y(x')|X = x]$. The observation of M_i can aid this inference because $M_i = M_i(x)$ for these units, and knowing $M_i(x)$ along with some background knowledge of how X affects M may provide information about $M_i(x')$. Knowing $M_i(x')$ can in turn provide information about $Y(x', M_i(x'))$ under some assumptions. A simple example will clarify this.

[2]Angrist et al. (1996) also assumes that the potential outcomes and potential treatments are well defined (i.e., the Stable Unit Treatment Value Assumption holds).

Consider the example presented in Section 1, where treatment ($X = 1$ if treated, $X = 0$ if not treated) is the presence of EDR, the mediating variable is voter registration ($M = 1$ if registered on election day, $M = 0$ if not registered on election day) and the outcome is turnout ($Y = 1$ if voted, and $Y = 0$ if did not vote). Consider the individuals that did not have EDR available ($X = 0$). For these individuals, we must make assumptions in order to make inferences about $\mathbb{E}[Y(1)|X = 0]$, which is unobserved. However, suppose we observe that a subset of these individuals registered but did not vote ($M = 1$ and $Y = 0$). Because one cannot vote without registering, we know that $Y_i(0, M_i(0)) = 0$ when $M_i(0) = 0$ and $Y_i(1, M_i(1)) = 0$ when $M_i(1) = 0$. If we are further willing to make an exclusion restriction analogous to the one made in Angrist et al. (1996) by assuming that $Y_i(0, M_i(0)) = Y_i(1, M_i(1))$ when $M_i(1) = M_i(0) = 1$, then it is straightforward to show that the observation of the a posttreatment variable can make these assumptions operational. If we observe someone who was not treated ($X_i = 0$) and did not vote ($Y_i = 0$) so that $Y_i(0) = 0$, then without observing the mediating variable, it is possible that they would have voted if treated ($Y_i(1) = 1$). However, if we observe that $M_i = 1$ so that $M_i(0) = 1$, then we know that this individual would not have voted if treated ($Y_i(1) = 0$) because $Y_i(1, M_i(1)) = Y_i(0, M_i(0)) = 0$ if $M_i(1) = 1$ due to the exclusion restriction, and $Y_i(1, M_i(1)) = 0$ when $M_i(1) = 0$ due to the registration requirement.

This demonstrates one possible way in which posttreatment variables can be used to provide information for causal inferences regarding total effects. There are many others (Pearl 1995, 2009; Joffe 2001; Tian and Pearl 2002a, 2002b; Kuroki and Cai 2008; VanderWeele 2008; Kaufman et al. 2009; VanderWeele and Robins 2010). In Section 3.2, we will discuss the implications of this particular approach within the context of turnout (and weaken some of these assumptions).

3 An Illustrative Example: The Effect of EDR on African American Turnout in Non-EDR States

Many studies (Rosenstone and Wolfinger 1978; Powell 1986; Nagler 1991; Highton 1997, 2004; Hanmer 2007; Achen 2008) utilize cross-sectional Current Population Survey (CPS) data at the individual level[3] to address the effects of relaxing registration laws (e.g., early deadlines for registration) on voter turnout.[4] As stated in the abstract of Rosenstone and Wolfinger (1978), the two key questions of such analyses are, "After the drastic relaxation of voter registration requirements in the 1960s, do present state laws keep people away from the polls? More specifically, which provisions have how much effect on what kinds of people?"

To demonstrate our methods, we address these questions by analyzing 2004 CPS data on African Americans—one of the most important "kinds of people" considered in this literature due to their history of disenfranchisement by state laws. We code the relevant variables as the following:

$$X \in \{0 \text{ (No EDR available)}, 1 \text{ (EDR available)}\}$$

$$Y \in \{0 \text{ (Did not vote)}, 1(\text{Voted})\}.$$

Table 1 presents the data on voting behavior on residence in an EDR state in a 2 × 2 table. In addition, we will, at some points adjust for a number of measured covariates including: family income, age, sex, and education. All these variables are as defined in the 2004 CPS.

3.1 *Traditional Approaches that Assume Conditional Ignorability of Treatment Assignment*

In this section, we look at two commonly used methods of making casual inferences using cross-sectional data—logistic regression and matching. The methods differ with respect to the particular functional form assumptions that are made, but each method makes essentially identical *causal* assumptions; namely, that

[3]The results from a weighted least squares analysis at the state level are substantively identical to those presented in this section.

[4]Unlike the other papers cited here, Achen (2008) compares turnout models that ignore registration with turnout models that utilize regression laws instrumentally to justify the identification of traditional effects in the turnout model (e.g., the effect of age on turnout). Furthermore, Achen (2008) examines a model for registration as a prelude to the joint modeling of registration and voting.

Table 1 Relationship between availability of election day registration and individual voting among African Americans. Each entry is the number of citizens in that category

	$Y = 0$ *Did not vote*	$Y = 1$ *Did vote*
$X = 0$ No EDR	1952	5170
$X = 1$ EDR	65	299

the counterfactuals are well defined and that treatment assignment is conditionally ignorable given a set of measured covariates. To be clear, we do not think the conditional ignorability assumption is very plausible in this setting. As such, we do not put much faith in the accuracy of the resulting estimates. The purpose of this section is not to provide plausible estimates of the EDR effect but rather to provide some sense of what typical researchers might infer about the effect of EDR on voting based on the CPS data under study. One can certainly do better than this. In Section 3.2, we show how one can make more plausible causal inferences from exactly the same data by utilizing more plausible causal assumptions.

Throughout Section 3, our causal estimand will be a type of CATE known as the ATE on the control units (ATC). Within the context of our voting application, ATC is formally defined as:

$$\text{ATC} \equiv \mathbb{E}[Y(1) - Y(0)|X = 0].$$

In words, ATC is simply the change in turnout we would have expected in current non-EDR states if EDR had been implemented in all these non-EDR states in 2004. This corresponds most closely to the Rosenstone and Wolfinger (1978) question of whether state laws "keep people away from the polls." It is also useful to note ATC is simply the fraction of helped ($Y(1) = 1, Y(0) = 0$) units among the untreated units minus the fraction of hurt ($Y(1) = 0, Y(0) = 1$) units among the untreated units. This means that the sample version of ATC can be written as:

$$\text{SATC} \equiv \frac{\#\{Y_i(1) = 1, Y_i(0) = 0, X_i = 0\} - \#\{Y_i(1) = 0, Y_i(0) = 1, X_i = 0\}}{\#\{X_i = 0\}}. \tag{1}$$

The population version of ATC can be written analogously. This will be especially relevant for the analysis in Section 3.2.

3.1.1 Logistic regression

It is common for researchers to apply a logistic regression or probit model to data similar to the CPS data in order to infer the causal effect of EDR on voting. The extent to which such an enterprise will be successful depends on the extent to which ignorability holds conditional on the covariates as well as how accurately the researchers' regression models approximate the true conditional expectation of voting given EDR status and the other covariates.

Although researchers typically only report coefficient estimates and their standard errors (or perhaps simple differences in fitted probabilities), these quantities will not typically correspond directly to ATC. Nonetheless, it is easy to use such regression results to construct an estimate of ATC (e.g., see Chapter 5 of Pearl 2009). Specifically, we can estimate ATC with

$$\widehat{\text{ATC}}_{\text{reg}} = \frac{1}{n_C} \sum_{i=1}^{n_C} \{\mu(X = 1, \mathbf{z}_i, \hat{\alpha}, \hat{\boldsymbol{\beta}}) - \mu(X = 0, \mathbf{z}_i, \hat{\alpha}, \hat{\boldsymbol{\beta}})\}, \tag{2}$$

where the $i = 1, \ldots, n_C$ index represents those individuals in non-EDR states ($X = 0$), $\mathbf{z}_i$ is the vector composed of individual i's observed covariates (family income, education, sex, age, and a constant term), and

$$\mathbb{E}[Y|X = x, \mathbf{z}] = \mu(X = x, \mathbf{z}, \alpha, \boldsymbol{\beta}) = \frac{\exp(x\alpha + \mathbf{z}'\boldsymbol{\beta})}{1 + \exp(x\alpha + \mathbf{z}'\boldsymbol{\beta})}$$

is the conditional probability of voting given EDR status and the measured covariates under the logistic regression model.[5] Note that in order to estimate $\hat{\alpha}$ and $\hat{\boldsymbol{\beta}}$, we must utilize observations with $X = 1$ in addition to the observations with $X = 0$ in the regression. However, because we are interested in the ATE for the $X = 0$ individuals (ATC), the estimator averages the differences in the estimated regression functions (the individual terms in Equation (2)) according to the empirical distribution of the covariates ($\mathbf{z}$) among the control units.[6]

Table 2 reports results from a series of nine logistic regression models. These models range from the fairly flexible model 1 in which all two-way interactions are present along with three-way interactions between (EDR, sex, family income), (EDR, sex, age), and (EDR, sex, education) to the extremely parsimonious model 9 that only includes EDR and a constant term. Looking across the row that provides the estimates of ATC, we see that these estimates are remarkably stable across the various specifications—ranging from 0.096 under the parsimonious model 9 to 0.133 under model 5.[7] In all cases, the lower endpoint of the 95% confidence interval does not fall below 0.056 and is typically closer to 0.09. Taken as a whole, these results would seem to suggest an effect of EDR on voting among non-EDR-state African American residents that is around a 10% point increase. Of course, all these estimates rely on the fairly implausible assumption that assignment to EDR is conditionally ignorable given the measured covariates.

3.1.2 Propensity score matching

Matching provides an alternative means to estimate the effect of EDR on voting among non-EDR-state residents that does not rely on the specific parametric assumptions of logistic regression.[8] We proceed by fitting a propensity score model[9] and then using the `GenMatch` function in the `Matching` package (Sekhon 2011) to create a matched data set. One-to-one matches were constructed to achieve balance on the estimated propensity scores as well as the observed covariates (family income, education, sex, and age).

Figure 1 provides a visual depiction of the pre- and postmatching conditional distributions of these variables given EDR status. Inspection of these figures suggests that balance on these variables was reasonable before matching and was improved by the matching procedure. The variable for which postmatching balance appears the worst is family income.

If we are satisfied with the degree of balance obtained by this procedure, we can use the matched data set to estimate ATC. Table 3 presents this estimate and associated measures of sampling variability. Here, we see that the matching procedure produces an estimate of ATC that is equal to 0.144 (a 14% point increase in turnout due to EDR among African Americans) with a 95% confidence interval from almost 0.09 to nearly 0.20.

3.2 *An Alternative Approach that Does Not Assume Conditional Ignorability*

As the results from the "traditional" analyses above appear to be quite stable, one might be tempted to infer that there is a large (9.5% points or more) effect of EDR on turnout among African American citizens living in non-EDR states. Of course, there are a number of reasons that we might question these

[5]It is possible to allow the logistic regression model to include interactions between the treatment variable and the background covariates. This creates no problems other than making the notation somewhat clumsy. In this case, it should be understood that $\mu(X = x, \mathbf{z}_i, \alpha, \boldsymbol{\beta})$ is formed by setting $X = x$ in both the main effect and all interactions in which X appears.

[6]If we instead wanted an estimate the ATE for all observations, we would average these differences according to the empirical distribution of the covariates for the control and the treated individuals.

[7]The change in effect size from specification appears to be the result of the changing set of cases (due to the use of list-wise deletion) that the empirical average is being taken over. In results, not reported here, we fit these same models to the data set consisting of the 6302 observations with complete data. The estimated value of ATC across these results is remarkably stable around 0.13.

[8]For recent work in political science using matching see, among others, Ho et al. (2007) and Sekhon (2008).

[9]The propensity score model for the EDR indicator variable with family income, age, education, and sex as the predictor variables was fit with thin-plate regression splines and the smoothing parameter was chosen with generalized cross validation. Specifically, this model was constructed by using the `mgcv` library in `R` to fit a binomial generalized additive model with the following formula: `EDR ~ sex + s(famincome, age, educ, by=as.factor(sex))`. This allows the probability of treatment to be an arbitrary smooth function of family income, age, and education that differs across males and females.

https://doi.org/10.1093/pan/mpr021 Published online by Cambridge University Press

Table 2 Logistic regression estimates of the effect of EDR on voting among African Americans. Entries without brackets are point estimates, entries in brackets are 95% confidence intervals. The row labeled ATC presents the estimate of the ATE on the control units along with the associated 95% confidence interval.

	Model 1	*Model 2*	*Model 3*	*Model 4*	*Model 5*	*Model 6*	*Model 7*	*Model 8*	*Model 9*
Intercept	−7.227	−7.642	−10.103	−10.047	−1.66	−6.959	−9.658	−11.441	0.974
	(−10.927, −3.573)	(−11.284, −4.043)	(−11.265, −8.957)	(−11.186, −8.923)	(−1.948, −1.374)	(−7.933, −5.994)	(−10.786, −8.544)	(−12.464, −10.432)	(0.922, 1.026)
EDR	−9.923	1.811	1.868	0.919	0.895	0.844	0.901	0.622	0.552
	(−30.247, 9.114)	(−4.374, 7.853)	(−4.273, 7.81)	(0.59, 1.268)	(0.572, 1.24)	(0.524, 1.187)	(0.573, 1.25)	(0.34, 0.917(	(0.286, 0.833)
Family income	0.168	0.163	0.081	0.081	0.13	0.084	0.072		
	(0.116, 0.22)	(0.112, 0.215)	(0.065, 0.096)	(0.066, 0.097)	(0.116, 0.144)	(0.069, 0.099)	(0.057, 0.088)		
Sex	−1.474	−1.201	0.43	0.419	0.5	0.434		0.319	
	(−3.774, 0.827)	(−3.458, 1.058)	(0.307, 0.553)	(0.298, 0.54)	(0.382, 0.619)	(0.316, 0.553)		(0.21, 0.428)	
Age	0.041	0.041	0.026	0.025	0.019		0.026	0.027	
	(0.029, 0.054)	(0.029, 0.054)	(0.022, 0.03)	(0.022, 0.029)	(0.015, 0.022)		(0.022, 0.029)	(0.023, 0.03)	
Education	0.112	0.124	0.221	0.22		0.168	0.228	0.275	
	(0.02, 0.205)	(0.033, 0.215)	(0.192, 0.25)	(0.192, 0.248)		(0.143, 0.194)	(0.2, 0.257)	(0.251, 0.299)	
EDR * family income	−0.065	0.023	0.026						
	(−0.352, 0.229)	(−0.063, 0.111)	(−0.06, 0.115)						
EDR * sex	7.492	−0.38	−0.304						
	(−4.564, 19.82)	(−1.084, 0.318)	(−1.009, 0.396)						
Family income * sex	−0.056	−0.053							
	(−0.088, −0.024)	(−0.084, −0.022)							
EDR * age	−0.017	−0.009	−0.01						
	(−0.089, 0.059)	(−0.03, 0.012)	(−0.031, 0.011)						
Sex * age	−0.01	−0.01							
	(−0.017, −0.002)	(−0.017, −0.002)							
EDR * education	0.325	−0.002	−0.007						
	(−0.164, 0.845)	(−0.156, 0.155)	(−0.158, 0.15)						
Sex * education	0.071	0.063							
	(0.013, 0.129)	(0.006, 0.12)							
EDR * family income * sex	0.062								
	(−0.115, 0.241)								
EDR * sex * age	0.004								
	(−0.04, 0.047)								
EDR * sex * education	−0.218								
	(−0.534, 0.091)								
ATC	0.129	0.129	0.129	0.132	0.133	0.127	0.131	0.097	0.096
	(0.09, 0.166)	(0.089, 0.168)	(0.088, 0.168)	(0.094, 0.168)	(0.098, 0.172)	(0.083, 0.165)	(0.092, 0.169)	(0.057, 0.136)	(0.056, 0.134)
N	6302	6302	6302	6302	6302	6379	6302	7390	7486

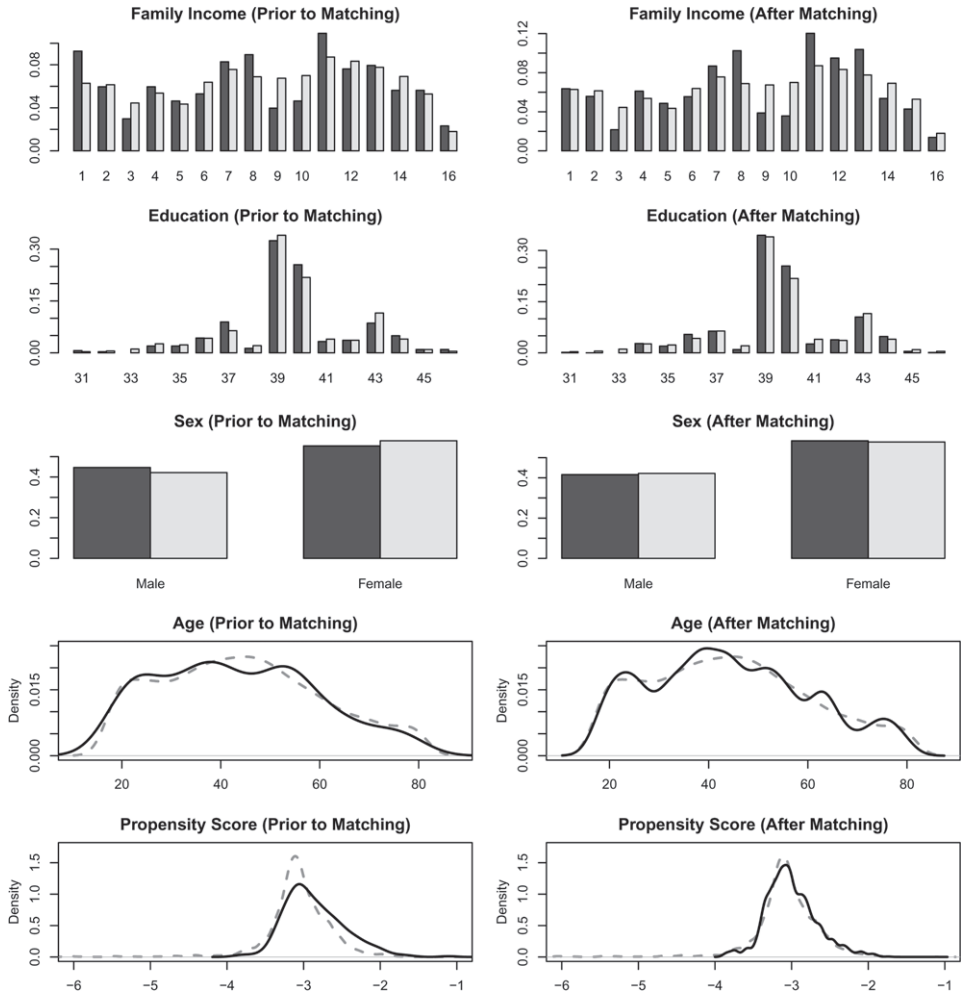

Fig. 1 Distribution of covariates conditional on EDR status before and after matching among African Americans. The dark bars in the barplots correspond to individuals in EDR states and the lighter bars correspond to individuals in non-EDR states. Similarly, the solid black lines in the density plots correspond to individuals in EDR states, whereas the dashed lines correspond to non-EDR state individuals.

seemingly robust results. There are far too few treated units to use as matches for control units (compare the two rows of Table 1), and there are undoubtedly unmeasured confounding variables that might affect the analysis or create overlap problems for measured confounding variables.[10] In particular, a sensitivity analysis that incorporates beliefs about unmeasured or unbalanced confounders might result in estimates that were smaller than 9.5%. However, we do not have to speculate to this extent because there are data *within the 2004 CPS* that can be utilized to demonstrate conclusively that the turnout effect cannot feasibly be as large as the results from the traditional analyses imply. In the remainder of this paper, we detail how one can more plausibly infer the effect on African American turnout of the adoption of EDR in non-EDR states using posttreatment variables from the same cross-sectional CPS data as the "traditional" analyses.

The basic idea underlying everything that follows is that by noting the equivalence of $Y_i(x)$ and $Y_i(x, M(x))$, we can break the numerator of Equation (1) into a number of pieces that depend on the

[10]For example, we might think that southern states are fundamentally different than northern states when it comes to the effects of registration laws on African American turnout. Unfortunately, no southern states had adopted EDR by 2004, so it is not possible to match southern non-EDR states with southern EDR states.

Table 3 Summary of effect of EDR on voting among those not in EDR states among African Americans-based on a matching analysis. The number of control (i.e., non-EDR) observations in the analysis is 6000 and the total unweighted number of observations is 6216

Estimand	*Estimate*	*SE*	*t-statistic*	95% *CI*
ATC	0.143	0.0274	5.229	(0.0896, 0.197)

observed values of M (registration) and Y (voting). Although *unconditional* assumptions about the fraction of helped and hurt units may be difficult to make and defend, assumptions that are conditional on M, Y (and X) are easier to make and defend. Conditioning on the posttreatment variable M thus allows one to make use of background information in a way that provides information about ATC without assuming that treatment assignment is (conditionally) ignorable.

In what follows we focus our attention on the upper bound for ATC. Similar reasoning could be used to calculate the lower bound for this quantity and, if desired, a fully Bayesian analysis could be conducted to construct point and interval estimates for ATC rather than just its bounds. The point of this section is not to provide a definitive estimate of the EDR effect. Instead, this section merely shows how conditioning on a posttreatment variable allows one to make very plausible assumptions about potential outcomes and that such assumptions result in an upper bound that is only above the traditional estimates of ATC in implausible circumstances and therefore shows the traditional estimates of ATC to be implausibly large.

Table 4 presents the data on the non-EDR individuals from Table 1 with a variable included for whether an individual was registered to vote. We note that this table contains a structural zero (as noted in Timpone 1998 and Achen 2008, you cannot vote without registering) and that the data are quite informative for the remaining three cells so that inference over the parameters of the observed variables is straightforward. We also note that EDR could not have had a positive effect on the 5170 individuals that voted, and as implied by the Manski bounds (Manski 2003), our estimate for the upper bound of ATC cannot be greater than $\frac{1276+676}{1276+676+5170} = 0.27$. In other words, we cannot estimate the effect of EDR to be more than 27% for this population of individuals. This upper bound is calculated in the same manner as the "worst-case" upper bound described in the Hanmer (2007) analysis of EDR for earlier years of the CPS.

However, when we observe the mediating variable of registration, we have reason to downgrade this upper bound even further. Suppose we are willing to assume that there would be a direct effect for only a small proportion of the 676 individuals that registered and did not vote. Formally, we can state this as the following:

Assumption 1. (Minimal direct effects among the registered). $Y_i(1, M_i(1)) - Y_i(0, M_i(0)) = 1$ when $M_i(1) = M_i(0) = 1$ for no more than a small proportion of the $X = 0$, $M = 1$, $Y = 0$ population, denoted with p_{dir}.

Under Assumption 1, our estimate for the upper bound of ATC cannot be greater than $\frac{1276+p_{\text{dir}}\cdot 676}{1276+676+5170}$. For example, if we assume that EDR would have had a direct effect on no more than 5% of the already

Table 4 Individual registration and individual voting among African Americans who did not have EDR available. Each entry is the number of citizens in that category in the 2004 CPS

	$X = 0$ No EDR	
	$Y = 0$ *Did not vote*	$Y = 1$ *Did vote*
$M = 0$ Did not register	1276	0
$M = 1$ Did register	676	5170

Table 5 Responses to the question: which of the following was the MAIN reason (you/name) (was/were) not registered to vote? The answers were provided by the 1276 non-EDR individuals who did not register or vote

U	*Main reason provided for not registering*	*Count*
1	Not interested in the election or politics	445
	My vote would not make a difference	46
	Not eligible to vote	94
0	Did not meet registration deadlines	209
	Did not know where or how to register	66
	Did not meet residency requirements/did not live here long enough	40
	Permanent illness or disability	94
	Difficulty with English	3
	Other reason	199
	Do not know	67
	Refused	12
	No response	1

registered individuals, then $p_{\text{dir}} = 0.05$ and our estimate for the upper bound of ATC cannot be greater than $\frac{1276+0.05\cdot 676}{1276+676+5170} = 0.18$. Note that this is an extremely conservative assumption because the most plausible story for a direct effect is based on the additional "advertising" that would accompany the adoption of EDR, and large randomized studies have found direct mail and phone encouragement effects of around 1% for African Americans (Green 2004).

Having reduced the estimated upper bound from the Manski bound of 27% to the new bound of 18%,[11] we can further reduce the bound by considering the 1276 individuals that did not register or vote. The first thing to note is that due to the registration requirement for voting, there can be no direct effects for these individuals. Formally, we state this as the following:

Assumption 2. (No direct effects among the unregistered). Due to registration laws, $Y_i(0, M_i(0)) = Y_i(1, M_i(1))$ when $M_i(1) = M_i(0) = 0$.

Under Assumption 2, for those with $M = 0$, the effect of the treatment on the outcome can only be positive when the effect of the treatment on the mediator is positive. Thus, we can further reduce the estimated upper bound for ATC by focusing solely on the effect of EDR on registration for 1276 individuals that did not register or vote. Fortunately, the 2004 CPS contains data that are relevant to exactly this question. Each of these 1276 respondents that did not register or vote was asked for the main reason they did not register. The counts of their responses are presented in Table 5. Although these responses do not provide definitive information about the percentage of the 1276 that would have registered if EDR had been available, the first three rows of $445 + 46 + 94 = 585$ responses clearly imply a lack of interest in politics, a lack of belief in the efficacy of voting, or an eligibility restriction, and hence we might think that these individuals would have been extremely unlikely to register in the case that EDR had been available. We define a binary variable U on the basis of these responses. For those uninterested individuals in the first three rows of Table 5, we say that $U = 1$. For the potentially interested individuals in rows 4–12 of Table 5, we say that $U = 0$.

Assumption 3. (Limited effects on the mediator among the unregistered and uninterested). $M_i(1) - M_i(0) = 1$ for at most a small proportion of the $U = 1, X = 0, M = 0, Y = 0$ population, who report themselves to be uninterested or ineligible. We let $p_{\text{reg}|U=1}$ denote this proportion.

[11] Hanmer (2007) employs alternative assumptions that do not utilize the measurement of a posttreatment variable in order to reduce the worstcase upper bound. For example, the upper bound can be reduced by assuming "capped outcomes" (that the average potential outcome under EDR in the non-EDR states will not exceed the average observed outcome for the EDR states). As noted in Hanmer (2007), this assumption will be controversial. We have intended our Assumption 1 (and the following Assumptions 2 and 3) to be relatively uncontroversial.

Under Assumptions 1–3, our estimate for the upper bound of ATC cannot be greater than $\frac{691+p_{\text{reg}|U=1}\cdot 585+p_{\text{dir}}\cdot 676}{1276+676+5170}$. For example, if we assume that EDR would have caused no more than 5% of the 585 uninterested/ineligible individuals to register, then $p_{\text{reg}|U=1} = 0.05$ and our estimate for the upper bound of ATC cannot be greater than $\frac{691+0.05\cdot 585+0.05\cdot 676}{1276+676+5170} = 0.11$. Again, note that this is a very conservative assumption. In order for the availability of EDR to have caused the $U = 1$ individuals to register, it would have had to interest them in voting as well as making it possible for them to register. Due to the aforementioned studies on advertising effects, it seems highly unlikely that EDR would have caused more than 5% of this group to register.

Note that under the very plausible Assumptions 1–3 with $p_{\text{dir}} = 0.05$ and $p_{\text{reg}|U=1} = 0.05$, our estimate for the upper bound of possible values for ATC is lower than the estimates we obtained using propensity score matching or using logistic regression models 1–7. This means that using a very simple analysis based on very plausible assumptions, we now know that most of the estimates we produced using the traditional analyses are *impossibly* large.

Furthermore, our estimate for the upper bound of ATC is only slightly higher than the smallest estimates of ATC that we achieved using the traditional approaches in the previous section (logistic regression models 8 and 9), and this is only because the upper bound implicitly assumes that all the 691 potentially interested $U = 0$ individuals from rows 4–12 of Table 5, who did not register or vote, will be caused to register and vote by the availability of EDR. It is straightforward to conduct a sensitivity analysis on these 691 individuals by defining the proportion that would have been caused to register and vote as $p_{\text{reg}|U=0}$. Using Assumptions 1–3 with $p_{\text{dir}} = 0.05$ and $p_{\text{reg}|U=1} = 0.05$, Fig. 2 presents the upper bound as a function of $p_{\text{reg}|U=0}$. The dashed lines represent 95% pointwise confidence intervals for the upper bound (which can be calculated in the standard asymptotic manner because the upper bound is a function of proportions). As a point of comparison, horizontal lines are presented to represent the estimates using matching and logistic regression (models 1–9) under conditional ignorability assumptions.

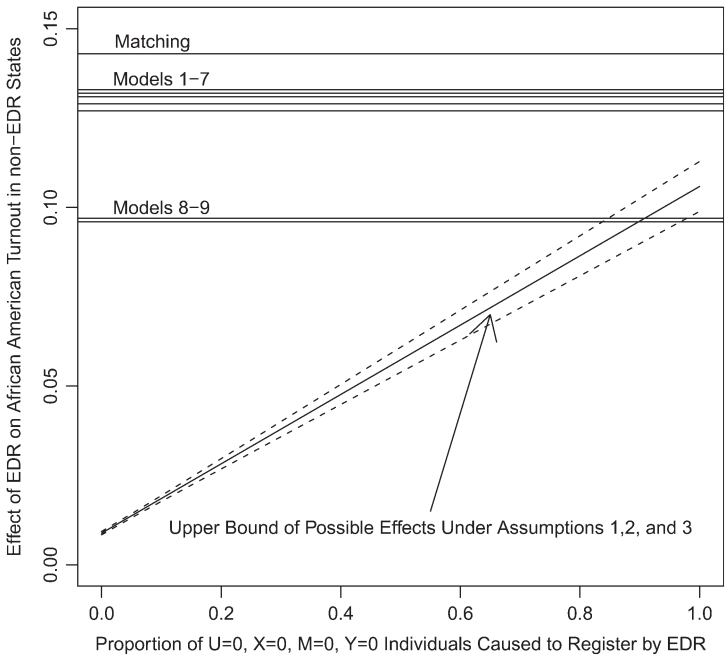

Fig. 2 Comparison of evidence on the effects of EDR on African American turnout in the non-EDR States (ATC). The upward sloping line represents the upper bound of possible effects under Assumptions 1, 2, and 3 with $p_{\text{dir}} = 0.05$ and $p_{\text{reg}|U=1} = 0.05$, and depending on what proportion of the unregistered ($M = 0$), nonvoting ($Y = 0$), but potentially interested ($U = 0$) population would be caused to register (and vote) by adoption of EDR ($p_{\text{reg}|U=0}$). The dashed lines represent 95% pointwise confidence intervals for the upper bound. The horizontal lines are presented for comparison and represent the estimates using matching and logistic regression (Models 1–9) under conditional ignorability assumptions.

There are two important things to note about this figure. First, as noted above, the matching estimate and the estimates from the logistic regression models 1–7 are all larger than the upper bound, regardless of the value for $p_{\text{reg}|U=0}$. Second, in order for an estimated upper bound on the possible values for ATC to be as large as 9.5%, we must have $p_{\text{reg}|U=0}$ at least as large as $0.89 = \frac{.095 \cdot 7122 - (0.05 \cdot 585 + 0.05 \cdot 676)}{691}$. In other words, if we believe that Assumptions 1–3 are plausible with $p_{\text{dir}} = 0.05$ and $p_{\text{reg}|U=1} = 0.05$, then in order to believe that the results from models 8 and 9 of the traditional analyses of the previous section are plausible, we must simultaneously believe that at least 89% of the unregistered individuals in rows 4–12 of Table 5 would have registered and voted had EDR been available to them. We suspect that many researchers would find this amount of voting and registration in a group that had not previously registered or voted to be implausibly high.

4 Discussion

Causal assumptions are a necessary part of making causal inferences from nonexperimental data. It is common for researchers to make such assumptions conditional on pretreatment covariates. The assumption that treatment assignment is conditionally ignorable given some set of measured pretreatment variables (also known as the "selection on observables" or "no unmeasured confounders" assumption) is frequently employed. It is extremely rare to see researchers use causal assumptions that are specified conditional on the value of a posttreatment variable. In this paper, we have shown how causal assumptions that are conditional on a posttreatment variable might, in some situations, be more plausible than the standard conditional ignorability of treatment assignment assumption. Further, in the example we studied, these more plausible assumptions would cause one to believe that standard estimates of the EDR effect are implausibly high.

Although there may be relatively few applications that are identical to the EDR example discussed in this paper, the general point of the paper is still valid and potentially useful in other applications. Our point is not that researchers should make exactly the same assumptions that we made here; rather, we hope to convince researchers that it is worthwhile to consider the possibility that subject matter expertise can be used to specify plausible causal assumptions conditional on values of posttreatment variables rather than just conditional on pretreatment variables.

References

Achen, C. H. 2008. Registration and voting under rational expectations: the econometric implications. Paper presented at the summer meeting of the society for political methodology, Ann Arbor, MI.

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. 1996. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* 91:444–55.

Balke, Alexander. 1995. Probabilistic counterfactuals: semantics, computation, and applications. PhD diss., University of California, Los Angeles.

Brady, Henry E., David Collier, and Jason Seawright. 2006. Toward a pluralistic vision of methodology. *Political Analysis* 14: 353–68.

Cochran, William G. 1965. The planning of observational studies of human populations (with Discussion). *Journal of the Royal Statistical Society Series A* 128:134–55.

Collier, David, and Henry E. Brady. 2004. *Rethinking social inquiry: Diverse tools, shared standards*. Lanham, MD: Rowman & Littlefield.

Copas, J. B. 1973. Randomization models for the matched and unmatched 2× 2 tables. *Biometrika* 60:467.

Deaton, Angus. 2009. Instruments of development: Randomization in the tropics, and the search for the elusive keys to economic development. NBER Working paper no. 14690.

Frangakis, C. E., and D. B. Rubin. 2002. Principal stratification in causal inference. *Biometrics* 58:21–9.

George, A. L., and A. Bennett. 2005. *Case studies and theory development in the social sciences*. Cambridge, MA: MIT Press.

Glynn, A. N. Forthcoming 2011. The product and difference fallacies for indirect effects. *American Journal of Political Science.*

Green, D. P. 2004. Mobilizing African-American voters using direct mail and commercial phone banks: A field experiment. *Political Research Quarterly* 57:245.

Hanmer, M. J. 2007. An alternative approach to estimating who is most likely to respond to changes in registration laws. *Political Behavior* 29(1):1–30.

Heckman, James, and Sergio Urzua. 2009. Comparing IV with structural models: What simple IV can and cannot identify. NBER Working paper no 14706.

Hedström, Peter. 2008. Studying mechanisms to strengthen causal inferences in quantitative research. In *The Oxford Handbook of Political Methodology*, eds. Janet Box-Steffensmeier, Henry E. Brady, and David Collier. Oxford: Oxford University Press.
Highton, B. 1997. Easy registration and voter turnout. *Journal of Politics* 59:565–75.
———. 2004. Voter registration and turnout in the United States. *Perspectives on Politics* 2:507–15.
Ho, Daniel E., Kosuke Imai, Gary King, and Elizabeth A. 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* 81:945–60.
Imai, Kosuke, Luke Keele, and Teppei Yamamoto. 2010. Identification, inference, and sensitivity analysis for causal mediation effects. *Statistical Science* 25:51–71.
Joffe, M. M. 2001. Using information on realized effects to determine prospective causal effects. *Journal of the Royal Statistical Society: Series B, Statistical Methodology* 759–74.
Kaufman, Sol, Jay S. Kaufman, and Richard F. MacLehose. 2009. Analytic bounds on causal risk differences in directed acyclic graphs with three observed binary variables. *Journal of Statistical Planning and Inference* 139:3473–87.
King, Gary. 1991. "Truth" is stranger than prediction, more questionable than causal inference. *American Journal of Political Science* 35:1047–53.
King, G., R. O. Keohane, and S. Verba. 1994. *Designing social inquiry: Scientific inference in qualitative research*. Princeton, NJ: Princeton Univ Press.
Kuroki, M., and Z. Cai. 2008. The evaluation of causal effects in studies with an unobserved exposure/outcome variable: Bounds and identification. Proceedings from the 2008 Conference of Uncertainty in Artificial Intelligence, Helsinki, Finland.
Manski, Charles F. 2003. *Partial identification of probability distributions*. New York: Springer.
Nagler, J. 1991. The effect of registration laws and education on US voter turnout. *The American Political Science Review* 85:1393–405.
Neyman, J. 1923. On the application of probability theory to agricultural experiments: Essay on principles, Section 9. (translated in 1990). *Statistical Science* 5:465–80.
Pearl, Judea. 1995. Causal diagrams for empirical research.*Biometrika* 82:669–710.
———. 2001. Direct and indirect effects. Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence, Seattle, WA. pp. 411–20.
———. 2009. *Causality: Models, reasoning, and inference*. 2nd ed. New York: Cambridge University Press.
———. 2010. The Mediation Formula: A guide to the assessment of causal pathways in non-linear models. Technical report R-363.
Powell Jr, G. B. 1986. American voter turnout in comparative perspective. *The American Political Science Review* 80:17–43.
Robins, J. M. 2003. Semantics of causal DAG models and the identification of direct and indirect effects. In *Highly Structured Stochastic Systems*, eds. Peter J. Green, Nils Lid Hjort, and Sylvia Richardson, 70–81. Oxford: Oxford University Press.
Robins, J. M., and S. Greenland. 1992. Identifiability and exchangeability for direct and indirect effects. *Epidemiology* 3:143–55.
Rosenbaum, Paul R. 1984. The consequences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society: Series A* 147:656–66.
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.
Rosenstone, S. J., and R. E. Wolfinger. 1978. The effect of registration laws on voter turnout. *The American Political Science Review* 72:22–45.
Rubin, Donald B. 1978. Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics* 6(1):34–58.
Sekhon, Jasjeet S. 2008. The Neyman-Rubin model of causal inference and estimation via matching methods. *The Oxford Handbook of Political Methodology*, ed. Janet M. Box-Steffensmeier, Henry E. Brady, and David Collier. New York: Oxford University Press.
Sekhon, Jasjeet Singh. 2011. Multivariate and propensity score matching software with automated balance optimization: the matching package for R. *Journal of Statistical Software* 42:1–52.
TenHave, Thomas R., Michael R. Elliot, Marshall Joffe, Elaine Zanutto, and Catherine Datto. 2004. Causal models for randomized physician encouragement trials in testing primary care depression. *Journal of the American Statistical Association* 99(465):16–25.
Tian, J., and J. Pearl. 2002a. A general identification condition for causal effects. Proceedings of the National Conference on Artificial Intelligence. Menlo Park, CA: AAAI Press/MIT Press. pp. 567–73.
———. 2002b. On the identification of causal effects. Proceedings of the American Association of Artificial Intelligence. Menlo Park, CA: AAAI Press/MIT Press.
Timpone, R. J. 1998. Structure, behavior, and voter turnout in the United States. *American Political Science Review* 92:145–58.
VanderWeele, T. J. 2008. The sign of the bias of unmeasured confounding. *Biometrics* 64:702–06.
VanderWeele, T. J., and J. M. Robins. 2010. Signed directed acyclic graphs for causal inference. *Journal Royal Statistical Society Series B* 72:111–27.