Explaining Causal Findings Without Bias: Detecting and Assessing Direct Effects

Faculty Research Working Paper Series

Avidit Acharya

Stanford University

Matthew Blackwell

Harvard University

Maya Sen

Harvard Kennedy School

October, 2015

RWP15-064

Explaining Causal Findings Without Bias: Detecting and Assessing Direct Effects*

Avidit Acharya,† Matthew Blackwell,‡ and Maya Sen§

October 19, 2015

Abstract

Researchers seeking to establish causal relationships frequently control for variables on the purported causal pathway, checking whether the original treatment effect then disappears. Unfortunately, this common approach may lead to biased estimates. In this paper, we show that the bias can be avoided by focusing on a quantity of interest called the controlled direct effect. Under certain conditions, the controlled direct effect enables researchers to rule out competing explanations—an important objective for political scientists. To estimate the controlled direct effect without bias, we describe an easy-to-implement estimation strategy from the biostatistics literature. We extend this approach by deriving a consistent variance estimator and demonstrating how to conduct a sensitivity analysis. Two examples—one on ethnic fractionalization's effect on civil war and one on the impact of historical plough use on contemporary female political participation—illustrate the framework and methodology.

*Thanks to Adam Cohon, Justin Esarey, Adam Glynn, Robin Harding, Gary King, Macartan Humphreys, Kosuke Imai, Bethany Lacina, Jacob Montgomery, Judea Pearl, Dustin Tingley, Teppei Yamamoto, and conference or workshop participants at Dartmouth, Harvard, Princeton, WashU, the Midwest Political Science Association meeting, and the Society for Political Methodology summer meeting for helpful discussions and comments. Thanks to Anton Strezhnev for valuable research assistance. Any remaining errors are our own.

†Assistant Professor of Political Science, Stanford University. email: avidit@stanford.edu, web: http://www.stanford.edu/~avidit.

‡Assistant Professor of Government, Harvard University. email: mblackwell@gov.harvard.edu, web: http://www.mattblackwell.org.

§Assistant Professor of Public Policy, Harvard University. email: maya_sen@hks.harvard.edu, web: http://scholar.harvard.edu/msen.

1 Introduction

Rigorous exploration of causal effects has become a key part of social science inquiry. No longer is it sufficient for researchers to claim a causal finding without additional theorizing and evidence of how the effect came to be. For many scholars, these inquiries usually involve ruling out different possible explanations. For example, does colonial status affect development even among countries with similar kinds of institutions? Does the incumbency advantage in American politics operate even when challengers are of comparable quality? Does a country's natural resource wealth affect democracy even among countries with identical levels of state repression? These important debates illustrate that disparate literatures share a similar concern: does a causal finding remain even after controlling for factors that are realized after, and possibly due to, the treatment in question?

An extremely common approach to these types of questions is to simply condition on (or "control for") posttreatment variables—i.e., variables on the causal path leading from the treatment of interest to the dependent variable. The goal here, implicitly or explicitly, is for authors to produce a causal effect estimate washed of an alternative explanation. As an illustration of just how widespread this approach is, we reviewed all empirical papers published in three of the top journals in political science from 2010 to 2015.[1] Of these papers, 40% *explicitly conditioned on a posttreatment variable in the analysis*; an additional 27% conditioned on a variable that *could plausibly be posttreatment*.[2] (Only 33% of papers clearly had no posttreatment variables included in their analyses.) Thus, we estimate that as many as *two-thirds of empirical papers in political science that make causal claims condition on posttreatment variables*. This is a figure that includes both observational and

[1] These journals were the *American Political Science Review*, the *American Journal of Political Science*, and *World Politics*. Removing articles that had no empirical content left us with 587 papers from January 2010 until either the second or third issue of 2015, depending on the journal. Of these, 92 were explicitly descriptive papers and 64 had no clearly stated main variable of interest. Removing these left us with 431 papers.

[2] These "likely posttreatment" variables were often coded this way because it was unclear exactly when the relevant variables were measured. Typically, this occurred when two variables were measured in the same year of a panel dataset.

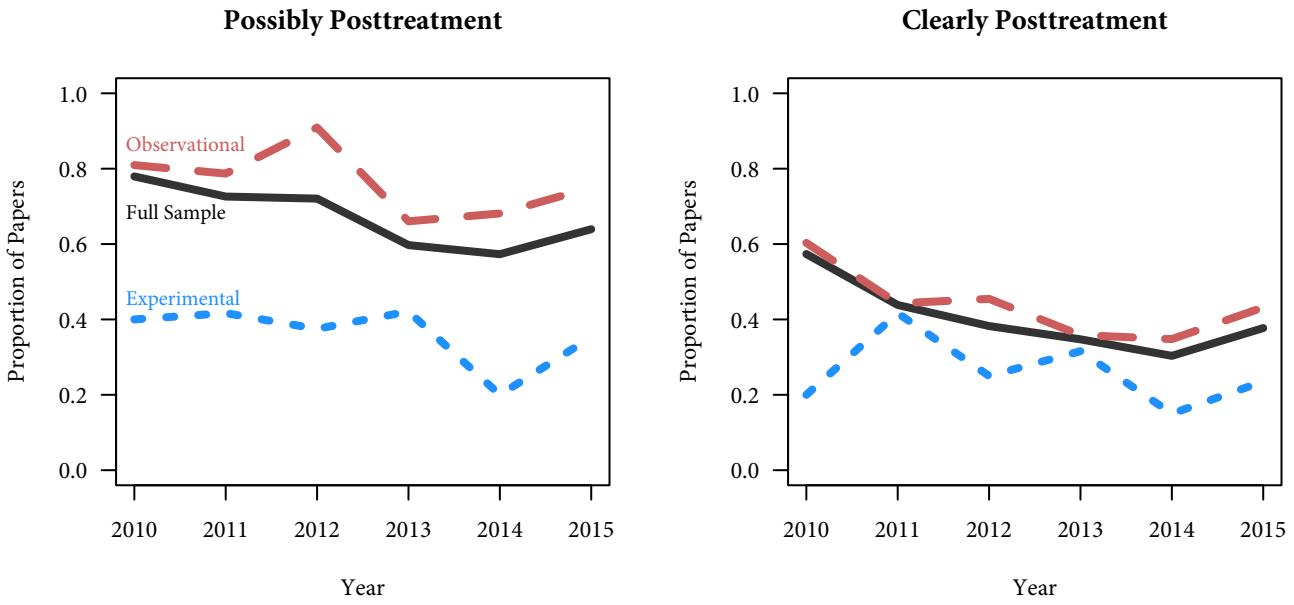


Figure 1: Proportion of papers implicitly or explicitly estimate causal relationships that include a posttreatment variable. Papers drawn from a systematic coding of all empirical articles from 2010 until 2015 in the *American Political Science Review*, the *American Journal of Political Science*, and *World Politics*. Left panel combines variables that are clearly posttreatment with those that are suspected to be posttreatment, but the original article is unclear on the measurement.

experimental studies (see Figure 1). Furthermore, when we reviewed why these variables were used, we discovered that about 23% of the papers that condition on posttreatment variables did so in order to explicitly test or adjudicate between causal mechanisms, with the rest attempting to control for alternative causal pathways. In short, this analysis suggests that a majority of empirical articles in political science are attempting to estimate the *direct effect* of a treatment fixing some consequences of that treatment.

Unfortunately, however, simply conditioning on posttreatment variables can result in seriously biased estimates of these direct effects. The key contribution of

this paper is to show that researchers can still estimate direct effects free of bias in a wide variety empirical settings. To do so, we focus on a simple quantity of interest: the *controlled direct effect* (CDE). The CDE represents *the causal effect of a treatment when the mediator is fixed at a particular level*. This allows researchers to rule out whether rival explanations or theories are the exclusive drivers of their findings, which, as our analysis of recent articles shows, is an important goal for many political scientists. In addition, the CDE is the quantity of interest identified from an experimental design where both the treatment and mediator are set to particular levels. Designs of this variety, such as traditional factorial designs and conjoint analyses, are playing an increasingly important role in our understanding of politics. While this makes the CDE well suited to answering policy and program evaluation questions, we show below that it can also speak to causal mechanisms under certain assumptions. Thus, the CDE is an important part of the applied researcher's causal toolkit.

A problem associated with the estimation of direct effects, controlled or otherwise, is what we call *intermediate variable bias*, which is attributable to *intermediate confounders*—or variables that are affected by the treatment and affect both the mediator and outcome. To estimate the controlled direct effect without such bias, we introduce a method from the biostatistics literature that addresses these challenges and is well-suited for continuous treatments and mediators. The method is implemented by way of a simple two-stage regression estimator, the *sequential g-estimator* (Vansteelandt, 2009; Joffe and Greene, 2009). This approach transforms (or *demediates*) the dependent variable by removing from it the effect of the mediator and estimates the effect of the treatment on this demediated outcome. Under certain assumptions, it has been shown that this is a consistent estimate of the controlled direct effect. We extend the usefulness of this approach by deriving a consistent variance estimator and describing how to conduct a sensitivity analysis for the key identification assumption. The methodology is easy to use, intuitive, and straightforward to implement with existing statistical software. We also provide open-source software that implements these methods and calculates the correct

variance estimates and conducts sensitivity analyses.

We note that the direct effect is part of a growing family of causal quantities that center on questions of direct and indirect effects. Another quantity, the natural direct effect (NDE), has been widely studied in both statistics (Pearl, 2001) and political science (Imai et al., 2011). The CDE and the NDE answer similar, yet distinct causal questions and each has different properties. For instance, the total causal effect (the average treatment effect) can be decomposed into two quantities: the natural indirect effect and the natural direct effect, making the NDE useful for evaluating causal mechanisms. As we show below the decomposition of the total effect using the controlled direct effect is more complicated. Below we discuss these differences and highlight how the CDE and NDE approaches complement one another, noting when each has advantages in applied work. Together, these quantities provide a robust way for researchers to investigate and explain their causal findings.

We proceed as follows. Section 2 explains why conditioning on posttreatment variables—the approach taken by many social scientists—has the potential to introduce serious bias. In Section 3, we explain direct effects and controlled direct effects, with Section 3.5 showing how controlled direct effects speak to causal mechanisms. Section 4 addresses the problems that result from the inclusion of intermediate confounders, while Section 5 presents sequential g-estimation (with our sensitivity analysis in Section 5.4). Section 6 illustrates both the framework and method via two empirical examples, which we use throughout for illustrative purposes. First, we replicate Fearon and Laitin (2003), to explore whether ethnic fractionalization affects civil war onset primarily via its impact on political instability. Using sequential g-estimation, we show that ethnic fractionalization has separate effect on civil war onset that does not operate through political instability. Second, we replicate Alesina, Giuliano and Nunn (2013), who show that historical plough use has an effect on modern-day female political participation. This effect exists only after conditioning on current-day income levels, which could introduce bias. We use sequential g-estimation to estimate the direct effect of the

plough not through income, finding an effect that is even stronger than in their original analyses. Section 7 addresses differences between the CDE and the NDE. We briefly conclude in Section 8 and provide additional technical material, including a discussion of our consistent variance estimator and replication `R` code, in the Appendix.

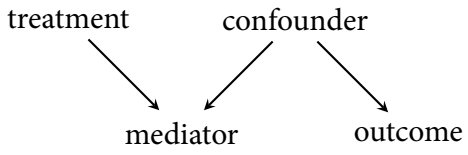
2 Why Conditioning on Posttreatment Variables Can Introduce Serious Bias

As we discussed above, up to two-thirds of articles within political science appear to condition on posttreatment (or possibly posttreatment) variables. Unfortunately, this practice raises two potentially serious complications. First, conditioning on a posttreatment variable changes the quantity of interest from an overall average treatment effect to a direct effect of the treatment *net the posttreatment variable*. This change may expected and may be the goal of these analyses—indeed, this is the quantity that we set out to estimate below—but it also might conflict with researchers' substantive interpretations. Even more sinister is the second consequence of conditioning on posttreatment variables, which is selection bias and which leaves the coefficient of interest without any causal interpretation. (This bias is part of what we call *intermediate variable bias*, below.) Selection bias of this kind occurs because the posttreatment variable could be endogenous and related to the outcome in non-causal ways. When this occurs, conditioning on this variable can induce spurious correlations between the treatment and the outcome (Rosenbaum, 1984).

A simple thought experiment helps to illustrate this form of selection bias. For many reasons, a reasonable assumption is that there exists no causal relationship between being in a car accident and having cancer; that is, being in a car accident does not plausibly lead someone to develop or be cured of cancer. However, if we were to condition on being a patient in a hospital (posttreatment to having been in a car accident), we would see a strong negative correlation between these two

conditions: patients are admitted to the hospital for either condition, so that if a patient is in the hospital and has *not* been in a car accident, they must be sick or unwell in some other way, increasing the likelihood that the patient has cancer. Thus, conditioning on being in a hospital would induce a negative relationship between being in a car accident and having cancer. However, this relationship does not exist because car accidents have a strong cancer-fighting effect, but rather because conditioning on being in the hospital produces additional information and breaks the statistical independence between car accidents and cancer.

A simple simulation shows this more concretely. We base the simulation on the following:



The "mediator" is the posttreatment variable in question. Neither the treatment nor the mediator have an effect (direct or indirect) on the outcome. The mediator is correlated with the outcome through their common cause, labelled as the confounder. We generate data from this process and plot the relationship between the treatment and both the confounder and the outcome. Figure 2 shows the results of one such draw. Looking at the overall relationships, including the grey and black dots, we see, as expected, that there is no effect of treatment on the confounder or the outcome. When we condition on the mediator, however, we see a much different picture. The black dots in Figure 2 are observations that have a value of the mediator in a certain range. Here, conditioning on certain values of the mediator induces strong correlations between the treatment and (1) the confounder (left plot) and (2) the outcome (right plot), even though there is no effect of the treatment on either by construction. This is selection bias that is induced by conditioning on the mediator. Indeed, as the statistical literature has long pointed out, this bias need not be conservative—here we are inducing relationships where none

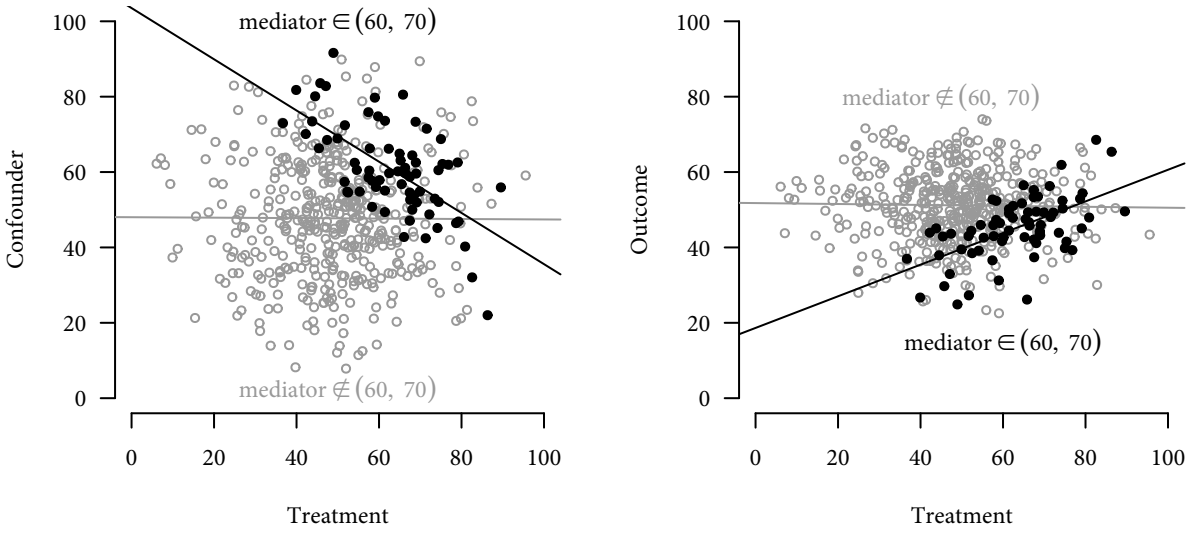


Figure 2: Simulated data showing the problem with conditioning on a posttreatment variable, here labelled "mediator." The black points and line are those when conditioning on the mediator being between 60 and 70 and the lighter grey lines are the regression lines for all points, black and grey. The figure on the left shows the relationship between the treatment and the confounder (none when all data are considered, but negative when conditioning on the mediator); the figure on the right shows the relationship between the treatment and the outcome (also none when all data are considered, but positive when conditioning on the mediator).

exist. Of course, when using real data, we will have no idea as to the direction of such bias.

What does this bias mean for the applied researcher? First, coefficients estimated from models that condition on posttreatment variables might not have a causal interpretation. Second, in some instances, the inclusion of posttreatment variables into a statistical analysis can bias the estimator *away from zero*, meaning that the estimates may overstate the size of a causal effect in either direction. Third, the bias means that any estimates may also be inconsistent: as sample sizes

increase, the coefficients from models that include posttreatment variables will actually converge to the wrong quantity. Thus, from the perspective of applied researchers, the problems associated with conditioning on posttreatment variables can be extremely serious.

3 What Are Direct Effects?

Given these problems, conditioning on posttreatment variables in a statistical model is seldom a suitable approach to causal inquiries. But how can researchers estimate direct effects without bias? In this section, we introduce a quantity of interest that fits the substantive goals of most researchers and can be estimated, free of bias, with a simple procedure. We start with the *direct effect* of a treatment, which is the effect of the treatment for a fixed value of the mediator (though how one "fixes" the mediator matters a great deal). Our goal in this paper is to estimate the *controlled direct effect* (CDE) of the treatment, which is the direct effect of the treatment when a mediator is the same fixed value *for all units*.[3] An *indirect effect*, on the other hand, is the portion of the total effect of treatment due to the treatment's effect on the mediator and the mediator's subsequent effect on the outcome.

We illustrate these concepts via a recurring example from comparative politics: Does ethnic fractionalization affect the onset of civil wars (Fearon and Laitin, 2003)? Specifically, does ethnic fractionalization increase the probability that a country will have a civil war primarily because it leads to greater political instability? Or does ethnic fractionalization increase the probability of civil war onset *independently* of greater instability? Our question of interest is whether ethnic fractionalization continues to have an independent effect on civil war onset holding political instability fixed at some value for all countries.

[3] Note that while the informal definition of the direct effect and our description of the CDE appear very similar, alternate definitions of direct effects exist and require different assumptions for estimation, as we discuss below.

3.1 Potential outcomes

To develop the framework, we rely on the potential outcomes model of causal inference (Rubin, 1974; Holland, 1986; Neyman, 1923), a counterfactual-based framework (that we summarize only briefly here). The advantage here, in contrast to a more traditional structural equation modeling approach, is that the potential outcomes framework allows us to incorporate heterogeneous causal effects easily and decouples the definition of a causal effect from its estimation. Let A_i be the treatment of interest, taking values $a \in \mathcal{A}$, where $\mathcal{A}$ is the set of possible treatment values. We define M_i as the mediator, taking values $m \in \mathcal{M}$. Throughout the paper, we assume both are continuous. Using our example, A_i represents ethnic fractionalization in a given country, while M_i represents the mediator, the country's political instability.

Studies of direct effects generally involve two sets of covariates. The first, which we call *pretreatment confounders* (X_i), are variables that affect the treatment and the outcome. The second, which we call *intermediate confounders* (Z_i), are those that are a consequence of the treatment and also affect the mediator. These are intermediate because they causally come between the treatment and mediator, as shown in Figure 3. For example, a country's average income is probably an intermediate confounder because it is (1) affected by ethnic fractionalization and (2) confounds the relationship between political instability and civil conflict. As we show below, such intermediate confounders cause major problems for the estimation of direct effects.

3.2 Total effects

Let Y_i be the observed outcome for unit i and $Y_i(a)$ be its *potential outcome*, the value that unit i would take if we set (and only set) the treatment to a. For instance, if the outcome was the country-level number of battle deaths due to civil conflict, then $Y_i(a)$ would be the number of battle deaths if ethnic fractionalization was set to a. The potential outcomes connect to the observed outcome by the *consistency*

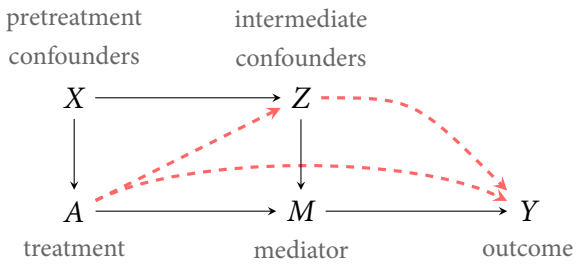


Figure 3: Directed acyclic graph showing the causal relationships present in analyzing causal mechanisms. Dashed red lines represent the controlled direct effect of the treatment not through the mediator. Unobserved errors are omitted.

assumption, $Y_i = Y_i(A_i)$, under which we observe the potential outcome for the observed treatment level.

In the potential outcomes framework, a causal effect is the difference between two potential outcomes, $\tau_i(a, a') = Y_i(a) - Y_i(a')$. This is the difference in outcomes if we were to switch unit i from treatment level a' to a. Under the consistency assumption, we only observe one of these potential outcomes for any unit—a problem known as the "fundamental problem of causal inference." To circumvent this, we typically estimate the average of treatment effects. We define the *average treatment effect* or *average total effect* (ATE, or τ) to be the difference in means between two different potential outcomes:

$$ATE(a, a') = E[Y_i(a) - Y_i(a')]. \tag{1}$$

where $E[\cdot]$ denotes the expectation over units in the population of interest. This is just the average effect if we were to change ethnic fractionalization from a' to a in all countries.

3.3 Controlled direct effects

To define controlled direct effects, we imagine intervening on both the treatment and the mediator at once. We define $Y_i(a, m)$ to be the value that the outcome that

unit i would take if we set its treatment to a and the mediator to m. For country-level number of battle deaths due to civil conflict, for example, $Y_i(a, m)$ would represent battle deaths if ethnic fractionalization was set to a and political instability to m. Under the same consistency assumption, then $Y_i = Y_i(A_i, M_i)$. The mediator also has potential values, $M_i(a)$, defined similarly to the potential outcomes; that is, $M_i(a)$ refers to the potential value that the mediator would take on under treatment level a. Applying the consistency assumption again, we have $M_i = M_i(A_i)$. Note that the potential outcome only setting a is the composition of these two potential outcomes: $Y_i(a) = Y_i(a, M_i(a))$.

The *controlled direct effect* (CDE) is the effect of changing the treatment while fixing the value of the mediator at some level m (Pearl, 2001; Robins, 2003):

$$CDE_i(a, a', m) = Y_i(a, m) - Y_i(a', m). \tag{2}$$

As with total effects, it is difficult to identify individual-level effects and so we focus on the average CDE or ACDE:

$$ACDE(a, a', m) = E[Y_i(a, m) - Y_i(a', m)]. \tag{3}$$

In other words, while a direct effect in general fixes the value of the mediator, the ACDE more closely corresponds to a *ceteris paribus* definition of a direct effect—that is, the direct effect with the mediator fixed at some value for all units in the population. It is the effect of separately fixing A_i and M_i to particular values for all units. In our example, this quantity of interest represents the effect of changing ethnic fractionalization if we were to fix the amount of political instability in a country at some level. The controlled direct effect is what is implicitly or explicitly estimated in experiments where units are randomized to receive more than one treatment, which are increasingly common in political science.

3.4 Natural direct and indirect effects

We note other important estimands for direct effects (Robins and Greenland, 1992; Pearl, 2001), which are common in the causal mediation literature. One quantity

of interest is the *natural direct effect* (NDE), which is the effect of changing treatment when fixing the mediator to its unit-specific potential value under a particular treatment level:

$$NDE_i(a, a') = Y_i(a, M_i(a')) - Y_i(a', M_i(a')). \tag{4}$$

The second value after the equality is simply the potential outcome $Y_i(a')$ under the treatment level a'. The natural direct effect represents the effect of a modified treatment that does not affect the mediator, but continues to directly affect the outcome. In our example this would be the effect of moving from complete ethnic homogeneity to some value a of ethnic fractionalization, holding political instability at what it would be under ethnic homogeneity. Note the crucial difference between the CDE and the NDE is at what level one "fixes" the value of the mediator.

A related quantity of interest is the *natural indirect effect* (NIE), which fixes the treatment and quantifies how the outcome changes only in response to treatment-induced changes in the mediator:

$$NIE_i(a, a') = Y_i(a, M_i(a)) - Y_i(a, M_i(a')). \tag{5}$$

The first term after the equality is, again, the potential outcome under a. The natural direct and indirect effect decompose the total effect for a single unit:

$$\tau_i(a, a') = NIE_i(a, a') + NDE_i(a, a'). \tag{6}$$

This decomposition also holds when we replace the individual-level quantities with their averages, $ATE(a, a') = ANIE(a, a') + ANDE(a, a')$, where ANIE and ANDE are the averages of the NIE and NDE, respectively.

We discuss the relative advantages of the ACDE and ANDE in Section 7.

3.5 How controlled direct effects speak to causal mechanisms

As we discuss below, the statistical literature on direct effects has mostly discussed the advantages of the ACDE in the context of experiments and policy evaluation.

While these goals are very useful for scholars, in this section, we show how the ACDE can also speak to causal mechanisms in two distinct ways. Note that the ANDE is the more straightforward estimand for evaluating mechanisms, but as we discuss below, it is often not identified in applied settings, leaving ACDE as a scholar's only option. Thus, it is important what information the ACDE can provide about mechanisms.

1) Ruling out alternative mechanisms. First, as VanderWeele (2011) notes, if the effect of a treatment is completely mediated by a mediator M_i and another set of potential mediators, W_i, then a non-zero ACDE for M_i implies that there must be an indirect effect that works through the set W_i.[4] Thus, showing that there is a non-zero ACDE implies that the effect of treatment is not due to the M_i mechanism *exclusively*. In Appendix B, we provide a formal proof of this result. For example, if our goal was to show that ethnic fractionalization had *some* effect on civil war deaths other than through political instability, we would estimate the ACDE and check its proximity to zero, taking into account uncertainty through a confidence interval or hypothesis test. A non-zero ACDE would suggest that there does exist an effect of ethnic fractionalization that does not operate exclusively through political instability. This is an intuitive approach for many applied researchers, who frequently want to rule out alternative explanations as being sole determinants of their findings.

2) Support for a preferred mechanism. Second, the difference between the ATE and ACDE summarizes the role of the mediator in a causal mechanism for the effect of A_i, allowing us to estimate support for a preferred mechanism. To see this, decomposing the total effect into three components is useful. VanderWeele (2014) and VanderWeele and Tchetgen Tchetgen (2014) show that, with a binary

[4] An effect is *completely mediated* by M_i and W_i if $Y_i(a, m, w)$ does not vary in a, where w denotes a fixed value for the tuple of mediators W_i.

mediator M_i, we can decompose the overall effect into the following:[5]

$$\tau(a, a') = \underbrace{ACDE(a, a', 0)}_{\text{direct effect}} + \underbrace{ANIE(a, a')}_{\text{indirect effect}} + \underbrace{E\big[M_i(a')[CDE_i(a, a', 1) - CDE_i(a, a', 0)]\big]}_{\text{interaction effect}}. \tag{7}$$

This decomposition shows that the between the ATE and ACDE is a combination of (1) the average natural indirect effect and (2) an interaction effect that captures how much the direct effect of A_i depends (causally) on M_i at the individual level. (Note that this interaction effect is distinct from the more typical "effect modification," which is a non-causal statement about how average effects vary as a function of potentially non-manipulated variables.) Under the following *constant interactions* assumption,

$$CDE_i(a, a', m) - CDE_i(a, a', m') = d(m - m'), \tag{8}$$

the interaction effect in (7) simplifies to $d \cdot E[M_i(a')]$ (whether the mediator is binary or continuous) and is identified under our assumptions in Section 5.1. Imai and Yamamoto (2013) explored this assumption to compare the indirect effect of two competing mechanisms. This result also implies that we can recenter M_i such that $E[M_i(a')] = 0$, and the total effect decomposes into the ACDE and the ANIE. Thus, under the constant interactions assumption and the identification assumptions below, the difference between the ATE and the ACDE is a measure of the indirect effect, provided we recenter M_i as $M_i^* = M_i - E[M_i(a')]$.

Without the constant interaction assumption, it will be infeasible to separate out the indirect effect from the interaction effect in (7). Even in this situation, the difference between the ATE and ACDE might still provide some information on how the causal finding came to be if we take a broader definition of causal mechanisms than is currently used in the statistics and biostatistics literature. Both the indirect effect and the interaction effect measure the impact of the mediator on how the treatment affects the outcome. The indirect effect measures how strong

[5] The decomposition can be easily generalized to the case of a continuous mediator. In that case the interaction effect is $\int_{m \in \mathcal{M}} E\big[CDE_i(a, a', m) - CDE_i(a, a', 0) | M_i(a') = m\big] dF_{M_i(a')}(m)$.

a particular causal pathway is, while the interaction effect tells us how much the mediator influences the direct effect of the treatment. The difference between the ATE and ACDE, then, will provide an aggregation of these two effects, which can be seen as a summary of how M_i participates in a causal mechanism for the effect of A_i on Y_i. This definition of a causal mechanism, however, is more expansive than previous definitions. For example, Imai, Keele and Yamamoto (2010) defines a causal mechanisms in terms of indirect effects. Thus, the difference in their framework between the ATE and the ANDE is the relevant measure of the strength of a mechanism. Referencing (7), this is equivalent to combining the ACDE and the interaction effect to create the ANDE. VanderWeele (2009), on the other hand, defines mechanisms in terms of the sufficient cause framework and shows that while there being an indirect effect through M_i implies that M_i participates in a mechanism, the reverse is not true. That is, there can be variables that participate in a mechanism that are unaffected by the treatment. For applied researchers, the empirical application will determine the utility of these approaches. In some cases, indirect effects are of particular interest while in others the combination of indirect effects and interaction is sufficient. We discuss these distinctions further in Section 7.

4 Intermediate Variable Bias

All direct effect quantities of interest—including both the ACDE and ANDE—raise the possibility of intermediate confounders. These intermediate confounders (Z_i in Figure 3) in turn raise the possibility of *intermediate variable bias*, a type of posttreatment bias. The intuition is as follows: conditioning on a mediator results in selection bias unless all of the intermediate confounders are included as well (sometimes called M-bias), but including them means including posttreatment variables (posttreatment bias). For instance, a researcher studying whether ethnic fractionalization affects civil war net political instability should be concerned whether there exist variables—measured or unmeasured—that (1) are affected by

ethnic fractionalization and also (2) affect political instability and civil war onset. Such variables are confounders for the mediator, but are also affected by treatment and thus are intermediate confounders. Unfortunately, dozens of such variables probably exist in our example—for example, religious fractionalization, average income, racism, and so on.

Most previous approaches to causal mediation confront this problem by *assuming that no intermediate confounders exist*. Unfortunately, this may be an unrealistic assumption for many researchers,[6] with the vast majority of mediators in the social sciences certainly violating this *no intermediate confounders* assumption. As Imai, Keele and Yamamoto (2010) point out, no intermediate confounders "is an important limitation since assuming the absence of post-treatment confounders may not be credible in many applied settings" (pg. 55). Moreover, *the ANDE and the ANIE are unidentified in the presence of intermediate confounders* without strong individual-level homogeneity assumptions (as we discuss in Section 5.1). This makes these quantitates of interest less attractive for most applied researchers.

The ACDE is, by contrast, identified in the face of intermediate confounders (also discussed in Section 5.1). Even so, the possible presence of intermediate confounders does raise the possibility of intermediate variable bias. To show this, we first assume a correctly specified linear model with constant treatment effects and no omitted variables for A_i. Under these assumptions, we can estimate the causal effect of A_i in a regression of the outcome on the treatment and the pretreatment confounders,

$$Y_i = \beta_0 + \beta_1 A_i + X_i^T \beta_2 + \varepsilon_i, \tag{9}$$

where β_1 is the (total) effect of A_i on Y_i. A common way to gauge the strength of some mechanism is to include a mediator, M_i, as an additional regressor in the model,

$$Y_i = \tilde{\beta}_0 + \tilde{\beta}_1 A_i + \tilde{\beta}_2 M_i + X_i^T \tilde{\beta}_3 + \varepsilon_i, \tag{10}$$

[6] Imai and Yamamoto (2013) present a method for causal mediation in the face of intermediate confounders, which they refer to as multiple causal mechanisms. This approach, however, requires these intermediate confounders to be themselves unconfounded, a similarly strong assumption.

and to interpret $\tilde{\beta}_1$ as a direct effect. Unfortunately, this interpretation is only correct under the assumption of no intermediate confounders. When these confounders, Z_i, are present, then $\tilde{\beta}_1$ will *not* equal the ACDE nor the ANDE, even under constant effects. This is because conditioning on a posttreatment variable can induce spurious relationships between the treatment and the intermediate confounders and, thus, the outcome (Rosenbaum, 1984).

In order to avoid this bias, we might decide to include the intermediate confounders in the regression:

$$Y_i = \alpha_0 + \alpha_1 A_i + \alpha_2 M_i + X_i^T \alpha_3 + Z_i^T \alpha_4 + \varepsilon_i. \tag{11}$$

Unfortunately, here too the coefficient on the treatment, α_1, will not be equal to the ACDE. This is because conditioning on Z_i blocks a possible causal pathway from $A \rightarrow Z \rightarrow Y$, which is part of the controlled direct effect. Thus, both omitting and conditioning on the intermediate confounders leads to a biased estimate of the ACDE.

5 Sequential g-estimation

In this section, we present sequential g-estimation, a method for estimating controlled direct effects in the face of intermediate confounders. Described by Vansteelandt (2009) and Joffe and Greene (2009), it is a type of structural nested mean model (Robins, 1986, 1994, 1997) that is tailored to estimating direct effects. We present (1) the assumptions that underlie this method, (2) basic identification results, (3) implementation details, and (4) an approach to sensitivity analysis.

5.1 Assumptions

As pointed out by Robins (1997), the ACDE is nonparametrically identified under what we call *sequential unconfoundedness*.

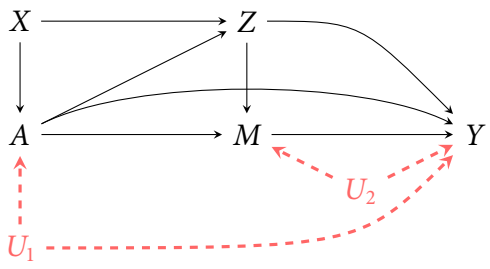


Figure 4: Directed acyclic graph showing a violation of sequential unconfoundedness with dashed lines representing omitted variable bias.

Assumption 1 (Sequential Unconfoundedness)**.**

$$\{Y_i(a,m), M_i(a)\} \perp\!\!\!\perp A_i | X_i = x \tag{12}$$

$$Y_i(a,m) \perp\!\!\!\perp M_i | A_i = a, X_i = x, Z_i = z, \tag{13}$$

for all possible treatment values $a \in \mathcal{A}$*, mediator values* $m \in \mathcal{M}$*, and covariate values* $x \in \mathcal{X}$ *and* $z \in \mathcal{Z}$*. In addition, we assume for all the above values:*

$$P(A_i = a | X_i = x) > 0 \tag{14}$$

$$P(M_i = m | A_i = a, X_i = x, Z_i = z) > 0. \tag{15}$$

This assumption represents two separate "no omitted variables" assumptions. First, no omitted variables for the effect of treatment on the outcome, conditional on the pretreatment confounders. Second, no omitted variables for the effect of the mediator on the outcome, conditional on the treatment, pretreatment confounders, and intermediate confounders. Thus, this represents a selection-on-the-observables assumption for each analysis. Because such assumptions can be unrealistic in observational studies, below we show how to weaken this assumption through a sensitivity analysis. Note that the pretreatment confounders can include pretreatment measurements of the intermediate confounders, which can add to the credibility of this assumption.

Figure 4 shows a situation where sequential unconfoundedness is violated. Here, dashed lines represent the effects of unmeasured confounders U_{i1} and U_{i2}. In the

Fearon and Laitin (2003) illustration below, U_{i1} represents the omitted variables for the effect of ethnic fractionalization on civil war onset and U_{i2} represents the omitted variables for the effect of political instability on civil war onset. Sequential unconfoundedness assumes that both paths $A_i \leftarrow U_{i1} \rightarrow Y_i$ and $M_i \leftarrow U_{i2} \rightarrow Y_i$ are absent, meaning that we have included enough variables in the pretreatment and intermediate covariates, X_i and Z_i, so that no omitted variable bias would be present in separately estimating the effect of either political instability or ethnic fractionalization on civil war onset.

In short, if the assumptions hold to separately estimate the effects of A_i and M_i on Y_i, then sequential unconfoundedness holds. Of course, this assumption might be violated if, for example, there are determinants of political instability and civil war onset that are not included in either X_i or Z_i. Because of the crucial nature of this assumption and the difficulty of showing that it holds in observational data, below we discuss a sensitivity analysis for assessing how large deviations from this assumption have to be in order to change the results of a study. Below, we apply this sensitivity analysis to the example of political instability and civil war onset.

Sequential unconfoundedness is not sufficient to identify the ANDE, however. To do so requires a stronger version of this assumption that omits intermediate confounders entirely so that both U_{i2} and Z_i must be absent from Figure 4 (Imai, Keele and Yamamoto, 2010). In addition, identifying the ANDE requires potential outcomes from different counterfactual worlds to be independent (Imai, Keele and Yamamoto, 2010) or for individual-level no- or constant-interaction assumptions (Robins, 2003; Imai and Yamamoto, 2013). Thus, one advantage of the ACDE is that it is identified under far weaker assumptions than the ANDE. Of course, the ANDE helps partition the total effect at a finer level and calculate the strength of a given causal pathway versus all other pathways. However, this additional information comes at the cost of these additional strong assumptions.

Even though ACDEs are identified under Assumption 1, the effects depend on the distribution of the intermediate confounders (Robins, 1997). To implement the simplest version of sequential g-estimation, we need the effect of the mediator

on the outcome to be independent of the intermediate confounders.

Assumption 2 (No Intermediate Interactions)**.**

$$\begin{aligned} &E\big[Y_i(a,m) - Y_i(a,m')\big|X_i = x, A_i = a, Z_i = z\big] \\ &\qquad = E\big[Y_i(a,m) - Y_i(a,m')\big|X_i = x, A_i = a\big], \end{aligned} \tag{16}$$

for all values $a \in \mathcal{A}$, $m, m' \in \mathcal{M}$, $z \in \mathcal{Z}$, *and* $x \in \mathcal{X}$.

This assumption has several notable features. First, this assumption is not required for the nonparametric identification of the ACDE. Nonparametric identification only relies on sequential unconfoundedness and this no intermediate interactions assumption only serves to make estimation simpler. In fact, we can relax this assumption to the extent that we are willing to model the distribution of Z_i conditional on A_i and X_i. Second, even if this assumption is false, the estimated effects will be weighted averages of ACDEs within levels of the intermediate confounders (Vansteelandt and Joffe, 2014, pp. 718). Thus, this assumption is similar to omitting an interaction term from a regression model. Third, this assumption does not rule out important interactions between the treatment and the mediator (see, for example, (24) below) nor interactions with the baseline covariates. Finally, this assumption is weaker than other no interaction assumptions (see, e.g. Robins, 2003) that restrict the controlled direct effect at the individual level or other assumptions that rule out intermediate confounders entirely.

5.2 Identification

In order to tune the estimator to estimating direct effects, it is useful to define the following function, which we call the *demediation function*[7]:

$$\gamma(a,m,x) = E[Y_i(a,m) - Y_i(a,0)|X_i = x] \tag{17}$$

[7] Demediation functions are commonly used in structural nested mean models in biostatistics where they are called blip-down functions (Robins, 1997).

This function is the effect of switching from some level of the mediator to o and does not depend on the value of the intermediate confounders due to Assumption 2. We call this the demediation function because when subtracted from the observed outcome, $Y_i - \gamma(A_i, M_i, X_i)$, it removes the variation in the outcome due to the causal effect of the mediator:

$$E[Y_i - \gamma(a, M_i, x)|A_i = a, X_i = x] = E[Y_i(a, 0)|X_i = x]. \tag{18}$$

This property of the demediation function follows easily from Assumptions 1 and 2 (Robins, 1994; Vansteelandt, 2009). Based on this, the ACDE conditional on X_i,

$$E[Y_i(a, 0) - Y_i(0, 0)|X_i = x],$$

is nonparametrically identified as difference in means of the demediated outcome:

$$E[Y_i - \gamma(a, M_i, x)|A_i = a, X_i = x] - E[Y_i - \gamma(0, M_i, x)|A_i = 0, X_i = x]. \tag{19}$$

The key intuition here is that after demediating the outcome, the remaining covariation with A_i is due to the direct effect of A_i.

Of course this result requires us to know the demediation function, which is unrealistic. Instead, we will almost always estimate it from data. Robins (1994) showed that, under sequential unconfoundedness, the causal difference γ is nonparametrically identified from the data and is equal to the difference-in-means estimator conditional on all the previous variables:

$$\widehat{\gamma}(a, m, x) = \widehat{E}[Y_i|A_i = a, M_i = m, X_i = x, Z_i = z] - \widehat{E}[Y_i|A_i = a, M_i = 0, X_i = x, Z_i = z]. \tag{20}$$

This follows from the simple fact that sequential unconfoundedness implies that the effect of the mediator on the outcome is identified. Furthermore, the identification of the ACDE as (19) holds when replace γ with its estimate, $\widehat{\gamma}$.

5.3 Implementation

When the treatment and mediator are binary or only take on a few values, nonparametric or semiparametric approaches exist to estimating the ACDE, reducing the

need for parametric models (Robins, Hernán and Brumback, 2000).[8] With a continuous treatment and a continuous mediator, nonparametric and semiparametric estimation of the difference-in-means in equations (20) and (19) have poor properties, including instability and high variability (Vansteelandt, 2009). Sequential g-estimation brings in parametric models to help estimate ACDEs in this context.

Sequential g-estimation proceeds in two simple steps. First, we regress the outcome on the mediator, treatment, and covariates (pretreatment and intermediate) to get an estimate of the demediation function. Second, we use the first stage to demediate the outcome and run a regression of this demediated outcome on the treatment and the pretreatment covariates. The marginal effect of the treatment in this second stage regression will be the estimate of the ACDE. We describe these steps in further detail.

5.3.1 First stage

The first stage of sequential g-estimation involves estimating the effect of M_i on Y_i, conditional on all other variables. The components of this model that involve M_i will be the parameterization of the demediation function, γ. For instance, we might use the following regression function:

$$E[Y_i|A_i, M_i, X_i, Z_i] = \alpha_0 + \alpha_1 A_i + \alpha_2 M_i + X_i^T \alpha_3 + Z_i^T \alpha_4. \quad (21)$$

This parametric model implies a parametric model on the demediation function, which is

$$\gamma(a, m, x; \alpha) = \gamma(m; \alpha_2) = \alpha_2 m. \quad (22)$$

We could augment this baseline regression model with interactions between the mediator and the treatment or the pretreatment confounders (but not with the intermediate confounders due to Assumption 2):

$$E[Y_i|A_i, M_i, X_i, Z_i] = \alpha_0 + \alpha_1 A_i + \alpha_2 M_i + X_i^T \alpha_3 + Z_i^T \alpha_4 + \alpha_5 M_i A_i + \alpha_6 M_i X_{ik}, \quad (23)$$

[8] See Blackwell (2013) for a political science application of such an approach.

where, X_{ik} is one variable in the matrix of pretreatment confounders. In this case, the model would imply a different demediation function:

$$\gamma(a, m, x; \alpha) = \alpha_2 m + \alpha_5 ma + \alpha_6 m x_k. \tag{24}$$

This more general demediation function allows the effect of mediator to vary by the values of the pretreatment confounders and the treatment.

The above specifications of the demediation function will generally identify the controlled direct effect with the mediator set to 0, $ACDE(a, a', 0)$. In some applications, this value of the mediator might be nonsensical or not of interest. In these cases and in cases where we are interested in exploring how the ACDE changes as a function of the mediator, it is possible to add another term to the demediation function to recenter the mediator:

$$\gamma(a, m, x, k; \alpha) = \gamma(m, k; \alpha_2) = \alpha_2(m - k). \tag{25}$$

When using this recentered demediation function, we will estimate $ACDE(a, a', k)$ in the second stage. We use this technique in the second empirical illustration below, where the mediator is log GDP of a country and a value of 0 is not a good point of comparison.

However we model the conditional mean of the outcome, we can obtain estimates of the parameters of the demediation function, $\hat{\alpha}$ from a least squares regression of the outcome on the treatment, mediator, pretreatment confounders, and intermediate confounders. Then, we can calculate the sample version of the demediation function,

$$\widehat{\gamma}(A_i, M_i, X_i; \widehat{\alpha}) = \widehat{\alpha}_2 M_i + \widehat{\alpha}_5 M_i A_i + \widehat{\alpha}_6 M_i X_{ik}. \tag{26}$$

The validity of this approach will depend on the validity of the modeling assumptions in (23). If this model for the conditional mean of the outcome is correct and Assumption 1 holds, then the least squares estimates will be unbiased for the parameters of the demediation function. To weaken the model dependence, one could extend this methodology to handle matching on pretreatment confounders.

With limited dependent variables, the conditional expectation function will not be truly linear so that there may be some bias in the estimation of the coefficients in (23) and thus the estimation of the demediation function. While linear specification can provide a poor approximation to the overall conditional expectation function of a nonlinear model, the linear model usually approximations marginal effects in these settings quite well (Angrist and Pischke, 2008). And since the relevant parameters of (23) and the demediation function are all in terms of these average differences, using a linear specification will likely result in relatively small amounts of bias in the estimation of the ACDE. Alternative models can be used to estimate the ACDE with non-continuous outcomes, but these typically require additional modeling and computation (Robins, 1997).

5.3.2 Second stage

With an estimate of the demediation function in hand, we can estimate the ACDE from the second stage model. First, we demediate the outcome,

$$\widetilde{Y}_i = Y_i - \widehat{\gamma}(A_i, M_i, X_i; \widehat{\alpha}), \tag{27}$$

which in our running example would be:

$$\widetilde{Y}_i = Y_i - \widehat{\alpha}_2 M_i - \widehat{\alpha}_5 M_i A_i - \widehat{\alpha}_6 M_i X_{ik}. \tag{28}$$

Given the results in Section 5.2, we can then estimate the ACDE of treatment by regressing this demediated outcome on the treatment and the pretreatment confounders,

$$E[\widetilde{Y}_i | A_i, X_i] = \beta_0 + \beta_1 A_i + X_i^T \beta_2, \tag{29}$$

where β_1 is the ACDE. The least squares estimator $\widehat{\beta}_1$ will be a consistent estimate of the ACDE and avoids any intermediate variable bias by not conditioning on either M_i or Z_i. Note that the standard errors on $\widehat{\beta}_1$ from this regression will be biased due to the fact that they ignore the first-estimation of γ. Note that similar caveats apply here in terms of limited dependent variables as in the first stage. In Appendix C we develop a consistent estimator for the variance of $\widehat{\beta}$ for linear models. One can also

use the nonparametric bootstrap, performing both stages of the estimation in each bootstrap replication. In simulations, these two approaches produced very similar results, though our variances estimator is far more computationally efficient.

Note that if Assumption 2 is violated, it is still possible to estimate the ACDE in a second stage, but that requires (i) a model for the distribution of the intermediate covariates conditional on the treatment and (ii) the evaluation of the average of within-stratum ACDEs across the distribution of that model. The second part entails a high-dimensional integral that is computationally challenging, though Monte Carlo procedure have been developed (Robins, 1986, 1997). Unfortunately, this generally requires additional modeling for the treatment which is unnecessary with sequential g-estimation. These approaches are mostly useful when large and interesting interactions are thought to exist in a particular context.

5.4 Sensitivity Analysis

To assess violations of sequential unconfoundedness, we provide a bias formula and sensitivity analysis in parametric models for the ACDE. Specifically, we derive the bias due to unmeasured intermediate confounders, which is a violation of (13) in sequential unconfoundedness. Take the following common LSEM structure as given:

$$Y_i = \alpha_0 + \alpha_1 A_i + \alpha_2 M_i + X_i^T \alpha_3 + Z_i^T \alpha_4 + \varepsilon_{iy}, \tag{30}$$

$$M_i = \delta_0 + \delta_1 A_i + X_i^T \delta_2 + Z_i^T \delta_3 + \varepsilon_{im}. \tag{31}$$

Under this parametric model, the sequential unconfoundedness assumption implies that ε_{iy} and ε_{im} are independent of one another. In the spirit of other approaches to sensitivity analysis (Imai, Keele and Yamamoto, 2010; Imbens, 2004; Blackwell, 2014) we can characterize the violation of this assumption with a parameter that measures the dependence between the errors of these two models:

$$\rho = \text{Cor}[\varepsilon_{iy}, \varepsilon_{im}]. \tag{32}$$

The sequential unconfoundedness assumption implies that $\rho = 0$. When $\rho \neq 0$, there are unmeasured covariates that affect both the mediator and the outcome, after controlling for A_i, X_i, and Z_i. By varying ρ, we can vary the severity of the unmeasured confounding for the effect of M_i. Here we make no assumptions about whether these unmeasured confounders are affected by treatment. We can characterize the bias in terms of this one parameter because there is a relationship between ρ and $\widetilde{\rho} \equiv \text{Cor}[\widetilde{\varepsilon}_{iy}, \varepsilon_{im}]$, where $\widetilde{\varepsilon}_{iy} \equiv Y_i - E[Y_i|A_i, X_i, Z_i]$.

To leverage this parameter, we can write the resulting bias of sequential g-estimation as a function of this error correlation. To do so, we adapt the approach to sensitivity analysis from Imai, Keele and Yamamoto (2010) to the context of the controlled direct effect. Given the LSEM structure of equations (30) and (31), suppose that (12) holds but that $\rho = \text{Cor}[\varepsilon_{iy}, \varepsilon_{im}] \neq 0$ so that (13) does not hold. The bias of the sequential g-estimate of the ACDE will be:

$$\text{plim}\, \widehat{ACDE}_{sg} - ACDE = -\widetilde{\delta}_1 \frac{\rho \widetilde{\sigma}_y}{\sigma_m} \sqrt{(1 - \widetilde{\rho}^2)/(1 - \rho^2)}, \tag{33}$$

where $\widetilde{\delta}_a = \partial E[M_i|A_i, X_i]/\partial A_i$, $\sigma^2_m = \text{Var}[\varepsilon_{im}]$, and $\widetilde{\sigma}^2_y = \text{Var}[\widetilde{\varepsilon}_{iy}]$. We show the derivation of this bias function in Appendix D. Furthermore, if ρ is known, the ACDE is identified. The identification here comes from the fact that the other parameters in the bias formula are identified on the remaining assumptions. Specifically, these parameters can be consistently estimated with the following additional regressions: (1) a regression of M_i on A_i, X_i, and Z_i; (2) a regression of M_i on A_i and X_i; and (3) a regression of A_i on X_i.

The bias formula (33) has several interesting features. First, the bias is 0 if either $\rho = 0$ or $\widetilde{\delta}_1$, the effect of the treatment on the mediator, is 0. Second, as is typical in such analyses (see, for example, Imai, Keele and Yamamoto, 2010), the bias is monotonic in the error correlation and its direction will depend on the sign of ρ. Third, this bias formula assumes no interaction between A_i and M_i, but it could be extended to handle such situations. Finally, we show in Appendix D that it is easy to reparameterize this sensitivity analysis in terms of the residual variation explained by the unmeasured confounder in the mediator and the outcome (Imbens, 2004;

Imai, Keele and Yamamoto, 2010; Blackwell, 2014).

6 Empirical Illustrations

We illustrate the controlled direct effect and sequential g-estimation via two empirical examples, one on the relationship between ethnic fractionalization, civil wars, and political instability (Fearon and Laitin, 2003) and the second on the long-term effects of plough-use on female participation in politics (Alesina, Giuliano and Nunn, 2013).

6.1 Ethnicity's effect on civil war onset

Fearon and Laitin (2003) show that the impact of ethnic fractionalization on the onset of civil wars fades when controlling for several posttreatment factors, including political instability and economic development. Here we use their general approach, but answer a more specific question in the spirit of the one of the key uses of the ACDE—to eliminate a rival mechanism. Specifically, we aim to address the potential theory that political instability is the sole mechanism driving any effect of ethnic fractionalization. The question is then whether any effect of fractionalization persists that does not operate through political instability. Our use of sequential g-estimation also avoids potential problems associated with posttreatment bias.

To estimate the ACDE of ethnic fractionalization net political instability, it is important to consider the timing of such a measurement. Fearon and Laitin (2003) define political instability to be any change of the Polity score by more than 3 in the last three years. Unfortunately for our current approach this makes the definition of the intermediate confounders more complicated. In order to stick to a more clear causal ordering, we define political instability as a change in Polity score of more than 3 between $t-2$ and $t-1$, where the outcome is measured at t. Keeping the original specification exactly the same as in Fearon and Laitin (2003) and applying sequential g-estimation results in extremely similar results.

With this definition of the mediator, we take the baseline set of covariates from Fearon and Laitin (2003) as the full set of possible pretreatment and intermediate confounders. We thus use their final model as a the first stage in our sequential g-estimation procedure. Next, we partition the covariates into pretreatment and intermediate, which requires some care in this case. First, note that it is not clear exactly what is pretreatment to ethnic fractionalization, but we use a heuristic that time-invariant variables are candidates for this category because they have the potential to have caused ethnic fractionalization in the past. Thus, we design the pretreatment covariates to include being a non-contiguous state, mountainous terrain, religious fractionalization, and being an oil exporter. We treat all other variables as intermediate confounders, including GDP per capita, log population, and Polity scores, all lagged by two years.[9] We diverge from the specification of Fearon and Laitin (2003) by lagging these variables to ensure that they are causally prior to our definition of political instability. Of course, it is possible that some of these confounders are miscategorized, such as religious fractionalization, which could be affected by ethnic fractionalization and thus would be an intermediate confounder. Good practice would be to combine different specifications of pretreatment/intermediate confounders and sensitivity analyses to ensure that results are not dependent on these choices.[10]

We then use the pretreatment variables only as controls in the second stage of the sequential g-estimation. For both stages, we use a linear probability model, with a binary variable for civil war onset as the dependent variable. We use the linear probability model because sequential g-estimation is valid for differences in means (or the *risk difference* in the parlance of epidemiology), but not for logistic models (Vansteelandt, 2010). A linear model in this context, without fully saturated covariates, can lead to bias from model misspecification, but such bias is usually low for estimating marginal effects which is the object of inference here

[9] Lagging by two years means that new state status, a covariate in the original Fearon and Laitin (2003) model, drops out of our specifications.

[10] Moving religious fractionalization to the intermediate confounders group has no effect on the results presented here. See replication materials for more details.

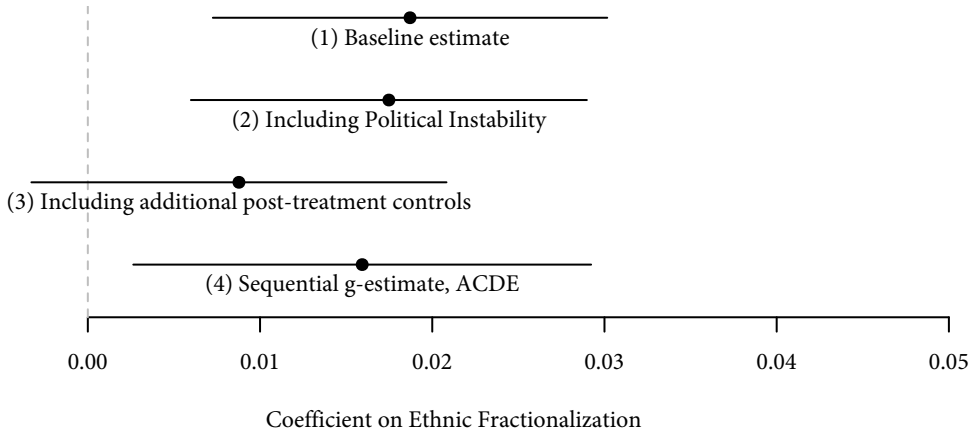


Figure 5: Estimated effects of ethnic fractionalization on civil war onset. Lines are 95% confidence intervals, with fourth model for ACDE using standard errors derived in Appendix C. Data from Fearon and Laitin (2003).

(Angrist and Pischke, 2008). Using the more complicated double logistic approach of Vansteelandt (2010), which should eliminate such issues, does not appear to substantively change the results. Standard errors come from the variance estimator in Appendix C and are similar to bootstrapped results.

Figure 5 shows the results of this analysis, along with (1) a baseline model with only pretreatment covariates and (2) a model that only includes political instability in addition to those covariates. Model (3) represent the effect of ethnicity as reported by Fearon and Laitin—it is sharply lower in magnitude and is not statistically significant. When we use sequential g-estimation in model (4), however, we find that the direct effect is almost identical to the original baseline estimate.[11] Thus, it appears a fairly strong direct effect of ethnic fractionalization exists even if all countries had no political instability. Given the discussion in Section 3.5,

[11] A Hausman-style test of the difference between models (3) and (4) reject the null that the two specifications are identical ($p < 0.001$).

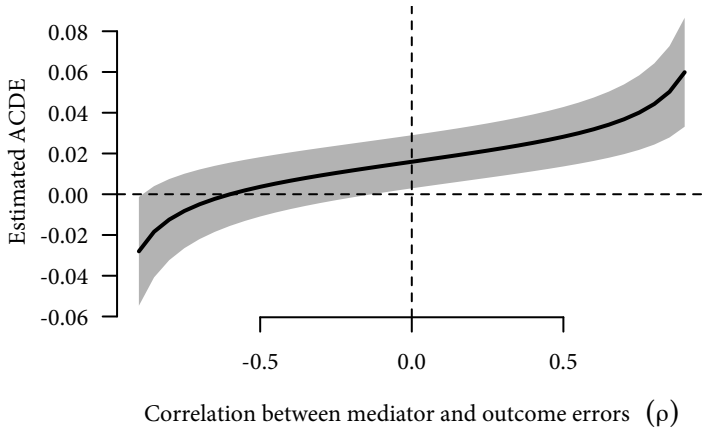


Figure 6: Sensitivity analysis for Fearon and Laitin (2003). Shaded regions are 95% confidence intervals.

a substantive interpretation is that some mediator other than political instability appears to have an indirect effect on civil war onset. In other words, political instability alone cannot explain the baseline results.

Our estimates might be biased if unmeasured confounders for the relationship between political instability and civil conflict exist. We therefore use the above sensitivity analysis and plot the results in Figure 6. The x-axis represents the residual, bias-inducing correlation between political instability and the onset of civil conflict after accounting for the observed baseline and intermediate confounders and the y-axis is the estimated ACDE under that amount of unmeasured confounding. With just a small negative residual correlation, the 95% confidence intervals (grey shaded regions) would overlap zero and indicate no significant ACDE. However, if political instability and civil conflict are positively associated (which is perhaps more plausible), then the ACDE will be greater than the original estimates. Either way, this sensitivity analysis allows us to determine what effect unmeasured confounding will have in any empirical application.

6.2 Effect of plough use on female political participation

Alesina, Giuliano and Nunn (2013) study the impact of a historical plough-based agriculture on modern-day female workforce and political participation. Using a measure of the relative proportion of ethnic groups that traditionally used the plough within a country (A_i, in our setup above), Alesina, Giuliano and Nunn (2013) regress various outcomes related to gender on this ploughs variable along with a set of historical controls. These historical controls represent the baseline controls, X_i from above. The core finding of Alesina, Giuliano and Nunn (2013) is that plough use has a strong, negative effect on female labor force participation and the share of firms owned by women, both measured in 2000. They argue that plough-based agriculture led to gender-based distinctions in labor specialization, with men tending to work in the fields (where their physical strength helped them use the plough) and women tending to work within the home. Over time, this specialization led to norms about the role of women in both the workforce and society more broadly, which were then passed down over generations until today, when we still see their effects in female labor force participation.

While Alesina, Giuliano and Nunn (2013) find strong effects for female labor force participation, they find smaller and statistically insignificant effects on the share of *political positions* held by women in 2000, which is our focus in this illustration (that is, this is our Y_i). However, after controlling for log GDP per capita in 2000 (which we take to be the mediator, M_i), a statistically significant coefficient for the ploughs variable emerges. Is this evidence of GDP being a key part of a mechanism for the ploughs effect? Alesina, Giuliano and Nunn believe so, and hypothesize that the null overall treatment effect of the plough is due to a positive indirect effect of the plough on incomes, and of incomes on female political participation. One issue with this particular choice of outcome and mediator is that there is some ambiguity about the causal ordering, given that they are measured in the same year. That is, it could be the case that female political participation, as measured by their representation in political positions, could affect GDP per capita. In spite of this potential problem, we hew as closely as possible to the orig-

inal analyses of Alesina, Giuliano and Nunn (2013) and use log GDP per capita in 2000 as our mediator.[12]

We use sequential g-estimation to evaluate this claim. When Alesina, Giuliano and Nunn (2013) control for current-day income in their regression model and then interpret the coefficient on ploughs as a direct effect, they implicitly assume no intermediate confounders, Z_i, for the effect of income on women's political participation. Given the historical time-frame, however, this assumption is implausible. To address this, we apply the sequential g-estimator with a set of intermediate controls designed to make the sequential unconfoundedness assumption more plausible.[13] In this context, the sequential unconfoundedness assumption essentially states that there are no omitted variables for the overall effect of the plough on female political participation and that there are no omitted variables for the effect of income on the same outcome, controlling for Z_i. In their empirical strategy, Alesina, Giuliano and Nunn (2013) include both log GDP and its squared term to account for non-linearity in the functional form. We take this approach in our development of the demediation function for this example:

$$\gamma(a, m, x; \alpha) = \alpha_2 m + \alpha_3 m^2 + \alpha_4 ma + \alpha_5 m^2 a. \tag{34}$$

Note that we also include an interaction term between the ploughs treatment and current-day income and income squared. Sequential g-estimation is flexible enough to handle complex empirical situations like this one. We recenter the log GDP variable (before squaring it, of course) here so that when we estimate the ACDE under $m = 0$, we are estimating the direct effect of the plough when log income is set to its mean. As discussion in Section 3.5, if the constant interaction assumption holds, then the difference between the ATE and ACDE under this recentering will be a measure of the strength of the mechanism.

[12] We replicated all results below using log GDP per capita in 2001 and the share of political positions held by women in 2010; all of the results are qualitatively similar.

[13] We include as intermediate controls civil conflict, years of interstate conflict, oil revenues per capita, proportion of population that is of European descent, a dummy for a former Communist country, the Polity score in 2000, and the value added of the services industry as share of GDP in 2000.

	Estimate	95% Bootstrapped CI
ATE	-2.10	[-6.00, 1.92]
Ignoring Z_i	-5.81	[-10.10, -2.30]
Sequential g	-7.87	[-13.21, -3.63]

Table 1: Estimates of the ATE and the ACDE of the plough on the percent of political positions held by women in 2000, where we show the naive and sequential g-estimates of the latter. For the ACDE, the mediator is log GDP per capita in 2000, set to its mean value. Nonparametric bootstrapped 95% confidence intervals based on 1,000 resamples shown in brackets.

In order to get a sense for how our estimates of a direct effect might differ given the assumptions that we make, we compare the overall effect of treatment to the estimated ACDE under the no intermediate confounders assumption and after accounting for intermediate confounders using sequential g-estimation. Table 1 presents these results and shows that while both approaches find a higher effect after accounting for current-day income, the estimated direct effect of the plough using sequential g-estimation is far stronger. How might this difference affect our inferences? To investigate this, we calculated the difference between the overall average treatment effect and the ACDE under each of these approaches. The distribution of the bootstrapped estimates for these measures of the strength of the causal mechanism of current-day income are shown in Figure 7. When ignoring the intermediate confounders, our conclusion would be that there is no statistically significant difference between the ATE and the ACDE (based on a 95% confidence interval) and, thus, income plays no statistically significant role in a causal mechanism. When we account for the intermediate confounders using sequential g-estimation, however, we find that all of the 1,000 bootstrapped estimates are above 0 and the average difference between the two estimates is much higher.

How can we interpret the result that the overall effect of ploughs is negligible, but the controlled direct effect appears to be strongly negative? Under the constant interaction assumption in Section 3.5 and our recentering of the mediator,

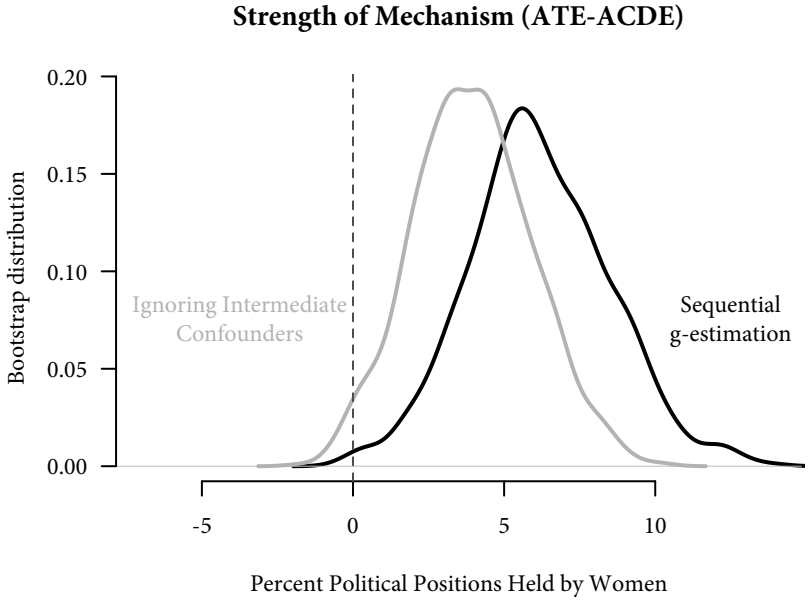


Figure 7: Bootstrap distributions of the difference between the ATE and ACDE, a measure of how strong a role income plays in the causal mechanisms of the use of ploughs is in Alesina, Giuliano and Nunn (2013). The smaller light grey distribution is ignoring intermediate confounders and the larger black distribution is the sequential g-estimation approach that accounts for intermediate confounders.

we can interpret this result as evidence for an indirect effect of ploughs through income. Without that assumption and given the decomposition in Section 3.5, we can only conclude that there is either some positive indirect effect of ploughs on female political participation through income or some causal interaction such that the effect of ploughs is weaker (more positive) in higher income areas. In either case, it is clear that national income is playing a role in producing the overall effect of ploughs on political participation. Again, we cannot determine how much of the effect is due to either of these sources without further assumptions, but we can use additional analyses to shed light on the matter. For instance, running an regression of income on ploughs and the historical covariates results in strong pos-

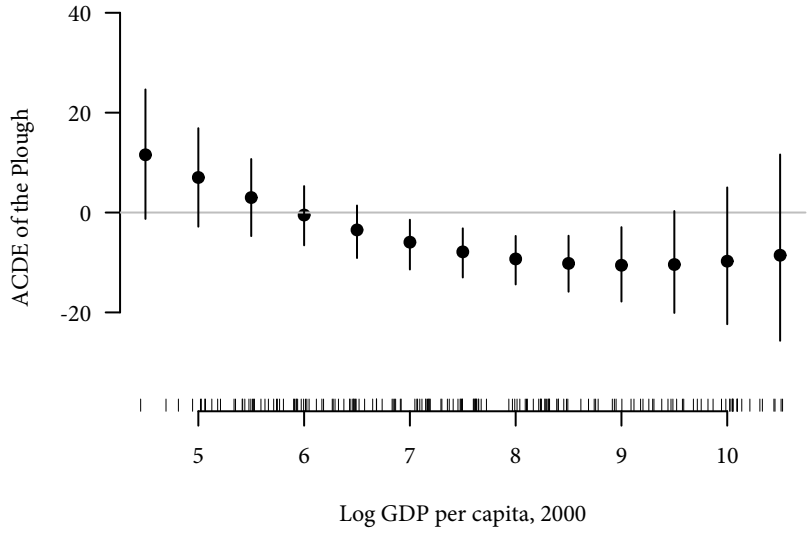


Figure 8: The ACDE of the plough as a function of the fixed level of current-day income. Vertical lines are 95% Confidence intervals from 1,000 bootstrapped replications.

itive effect of the plough on income. This appears to indicate that if we believed the additional assumptions of a traditional mediation analysis, we may conclude that there is an indirect effect of the plough through income.

In addition, we could investigate the causal interaction side of this analysis and find the $ACDE(m)$ for different levels, m, of current-day income. We do this by simply recentering M_i to different values and re-running the sequential g-estimation. We present the results in Figure 8, which shows that there does appear to be a (nonlinear) causal interaction between the plough and income. This interaction is negative, however, with richer countries having a more negative effect of the plough—the opposite of our above analysis. There are a number of possible reasons for this negative interaction but one might be the effect of economic development on gender norms either today or historically. That is, gender

norms encouraged by the plough could have been reinforced by subsequent economic growth or alternative gender norms could have developed in response to low levels of economic development. In the latter case, technological development in low-income, agricultural societies could induce moves away from the plough and, subsequently, changes in gender norms about work. In any case, these results are suggestive evidence that income plays an important role as part of both an indirect effect and an interaction with regard to the effect of the plough.

Finally, we note that Alesina, Giuliano and Nunn (2013) is part of a broader recent trend in social science exploring the long-term impact of historical institutions on contemporary factors (Banerjee and Iyer, 2005; Dell, 2010; Nunn and Wantchekon, 2011). As in Alesina, Giuliano and Nunn (2013), this brand of historical political economy by its nature must deal with mechanisms and persistence of the sometimes long-abolished institutions. Furthermore, the long historical gaps between when the treatment occurred (plough use, forced labor regimes, the slave trade) and the often contemporary mediators means that intermediate confounders are almost certainly an issue in these studies. The sequential g-estimation framework here allows for these types of studies to address the assumptions necessary to rule out alternative mechanisms and to actually estimate direct effects if those assumptions are met.

7 The ACDE and ANDE in Applications

When will the controlled direct effect be more appropriate for a given application than the natural direct effect, and vice versa? The literature is mixed. Pearl (2001) describes the ANDE as being useful for *descriptive* accounts and the ACDE as being useful for *prescriptive* accounts. For Vansteelandt and VanderWeele (2009), while "controlled direct effects are often of greater interest to policy evaluation (Pearl, 2001; Robins, 2003), natural direct and indirect effects may be of greater interest in evaluating the action of various mechanisms (Robins, 2003; Joffe, Small and Hsu, 2007)" (p. 459). Robins (2003) commented that the natural indirect effect,

"although possibly of mechanistic interest, may never be of direct public health interest, except as an approximation" (p. 10). Many papers (e.g., Imai, Keele and Yamamoto, 2010) take this reasoning a step further by equating the natural indirect effect with a mechanism. Rubin (2004), on the other hand, questions the entire approach, writing that "the concepts of direct and indirect causal effects are generally ill-defined and often more deceptive than helpful" (p. 162).

For the applied researcher, though, both the ACDE and ANDE can speak to causal mechanisms, albeit for slightly different definitions of the concept. We make several observations. First, stating the implied counterfactual comparison of each estimand is helpful. The ACDE counterfactual is *"what would the average effect of treatment be if we were to force the mediator be* m *for all units in the population?"* while the ANDE counterfactual is *"what would the average effect of treatment be if we forced every unit to take the value of the mediator it would have taken with* no *treatment?"* Thus, the ANDE estimates the effect of a modified treatment that has no effect on the mediator (because the mediator must be fixed at its "no treatment" value for each unit). Some questions better lend themselves to the former and others the latter, particularly when considering the hypothetical experimental analogy. For example, in considering the impact of ethnic fractionalization on civil war, a reasonable (if obviously hypothetical) experiment would be to intervene on political institutions to change the effect of fractionalization. In this case, the ACDE would be the appropriate estimand. However, the ANDE may be better suited in other studies. For example, in many framing studies within American politics, the treatment is usually an article with a certain frame and the mediator is often an emotional response to this treatment. Because the mediator is manipulable mostly through the treatment, the ANDE may be a better quantity of interest for these sorts of studies.

Second, in the same vein, the ACDE and ANDE correspond to different experimental designs. The ACDE is the quantity that is identified under a design with multiple treatments that are jointly randomized to each unit, such as in a 2 $\times$ 2 factorial design or a more complicated conjoint analysis. The ANDE is iden-

tified from more complicated experimental designs. Imai, Tingley and Yamamoto (2013) show that under a parallel design where some respondents have only their treatment status randomly assigned and others have their treatment and mediator randomly assigned, it is only possible to point identify the ANDE by assuming an individual-level no-interactions assumptions that implies the ACDE and ANDE are exactly the same. They also show that a cross-over design, where each unit is randomized at two different time periods (one for just the treatment and one for both treatment and mediator) requires an individual-level no carry-over assumption. Thus, the ideal experiment that identifies the ANDE will, in general, be more complicated and rely on assumptions beyond randomization as compared to the an experimental design for the ACDE. Of course, these additional assumptions do allow a more powerful identification result—under these designs it is possible to estimate the ANIE as well.

Third, the ANDE, due to its nested nature, combines (1) the direct effect (fixing the mediator at a particular value for all units) with (2) an interaction (allowing different values of the mediator for each unit in the population). The ACDE, however, omits this second component, evaluating the effect of intervening on both the treatment *and* the mediator. (Note that this means that the two estimands will be similar when there is little variation in the baseline or natural level of the mediator.) For the applied researcher, this means that a key question is whether the interaction, which includes the natural variation in $M_i(a)$ across units, is something that substantively should be included in the direct effect of the treatment. For example, in the ethnic fractionalization case, should the direct effect of ethnic fractionalization also include the interaction between a country's ethnic fractionalization and political instability, or should it not? Is such an interaction theoretically meaningful? These are substantive questions for researchers to consider.

Both the ANDE and the ACDE will be of use to the applied researcher. As we have seen, they speak to an overlapping set of causal questions, with the ACDE being estimable in a wider variety of contexts, but with the ANDE providing more information about the decomposition of the total effect. The causal questions un-

der investigation and the research design will often determine which of the two estimands is more appropriate in a given context.

8 Concluding Remarks

The rigorous exploration of causal effects is an essential component to social science research, and social scientists need a diverse set of tools to investigate competing theories and explanations. Unfortunately, as we have discussed, the current approach of many social scientists is to simply condition on posttreatment variables—an approach that has the potential to introduce serious bias. In light of this, the contributions of this paper are threefold. First, we called attention to an underused quantity of interest, the controlled direct effect—or the treatment effect holding "fixed" values of a potential mediator. Under certain assumptions, this quantity can help rule out alternative causal pathways and can detect whether a mediator participates in a causal mechanism. For many applied researchers, the ability to rule out alternative mechanisms is an essential part of a compelling empirical exploration. Second, we demonstrated how the usual approaches to estimating direct effects break down in the face of a common feature of social science applications: intermediate confounders.

Third, we presented and expanded a methodology for estimating controlled direct effects, sequential g-estimation, giving researchers the tools to use the method in applied work. This method has several properties that make it attractive for social scientists. First, the method estimates the controlled direct effect, which, as we argued, is of particular value to social scientists. Second, the method avoids the problems such as intermediate variable bias, which plague other methods. Third, sequential g-estimation relies on weaker assumptions than other methods for estimating direct effects. Lastly, the method is intuitive, straightforward, and easy to implement.

There are many areas of research that stand to benefit from the use of this approach to direct effects and causal mechanisms. As noted, the new trend in po-

litical economy studying the persistent effect of historical institutions engage with these questions on a regular basis. Similarly, a growing literature within international relations and comparative politics has examined the causal impact of complicated treatments, which rely exclusively on observational data. As these treatments could have a variety of effects, researchers working in these areas must address potential threats to their preferred explanation seriously. Our methodology provides these scholars with an important tool for evaluating such possibilities.

Bibliography

Alesina, Alberto, Paola Giuliano and Nathan Nunn. 2013. "On the origins of gender roles: Women and the plough." *Quarterly Journal of Economics* 128(2):469–530.

Angrist, Joshua D. and Jörn-Steffen Pischke. 2008. *Mostly Harmless Econometrics. An Empiricist's Companion* Princeton University Press.

Banerjee, Abhijit and Lakshmi Iyer. 2005. "History, Institutions, and Economic Performance: The Legacy of Colonial Land Tenure Systems in India." *American Economic Review* 95(4):1190–1213.

Blackwell, Matthew. 2013. "A Framwork for Dynamic Causal Inference in Political Science." *American Journal of Political Science* 57(2):504–520.
URL: http://www.mattblackwell.org/files/papers/dynci.pdf

Blackwell, Matthew. 2014. "A Selection Bias Approach to Sensitivity Analysis for Causal Effects." *Political Analysis* 22(2):169–182.

Dell, Melissa. 2010. "The Persistent Effects of Peru's Mining *Mita*." *Econometrica* 78(6):1863–1903.

Fearon, James D. and David D. Laitin. 2003. "Ethnicity, Insurgency, and Civil War." *American Political Science Review* 97(01):75–90.

Holland, Paul W. 1986. "Statistics and causal inference." *Journal of the American Statistical Association* 81(396):945–960.
URL: http://www.jstor.org/stable/2289064

Imai, Kosuke, Dustin Tingley and Teppei Yamamoto. 2013. "Experimental designs for identifying causal mechanisms." *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 176(1):5–51.

Imai, Kosuke, Luke Keele, Dustin Tingley and Teppei Yamamoto. 2011. "Unpacking the Black Box of Causality: Learning about Causal Mechanisms from Experimental and Observational Studies." *American Political Science Review* 105(04):765–789.

Imai, Kosuke, Luke Keele and Teppei Yamamoto. 2010. "Identification, Inference and Sensitivity Analysis for Causal Mediation Effects." *Statistical Science* 25(1):51–71.
URL: http://projecteuclid.org/euclid.ss/1280841733

Imai, Kosuke and Teppei Yamamoto. 2013. "Identification and Sensitivity Analysis for Multiple Causal Mechanisms: Revisiting Evidence from Framing Experiments." *Political Analysis* 21(2):141–171.
URL: http://pan.oxfordjournals.org/content/21/2/141.abstract

Imbens, Guido W. 2004. "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review." *Review of Economics and Statistics* 86(1):4–29.

Joffe, Marshall M, Dylan Small and Chi-Yuan Hsu. 2007. "Defining and Estimating Intervention Effects for Groups that will Develop an Auxiliary Outcome." *Statistical Science* 22(1):74–97.

Joffe, Marshall M and Tom Greene. 2009. "Related causal frameworks for surrogate outcomes." *Biometrics* 65(2):530–538.

Newey, Whitney K. and Daniel McFadden. 1994. Large sample estimation and hypothesis testing. In *Handbook of Econometrics*, ed. Robert F. Engle and Daniel L. McFadden. Elsevier pp. 2111–2245.

Neyman, Jerzy. 1923. "On the application of probability theory to agricultural experiments. Essay on Principles. Section 9." *Statistical Science* 5:465–480. Translated in 1990, with discussion.

Nunn, Nathan and Leonard Wantchekon. 2011. "The Slave Trade and the Origins of Mistrust in Africa." *American Economic Review* 101(7):3221–3252.

Pearl, Judea. 2001. Direct and Indirect Effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*. UAI'01 San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. pp. 411–420.
URL: http://dl.acm.org/citation.cfm?id=2074022.2074073

Robins, James M. 1986. "A new approach to causal inference in mortality studies with sustained exposure periods-Application to control of the healthy worker survivor effect." *Mathematical Modelling* 7(9-12):1393–1512.
URL: http://biosun1.harvard.edu/~robins/new-approach.pdf

Robins, James M. 1994. "Correcting for non-compliance in randomized trials using structural nested mean models." *Communications in Statistics* 23(8):2379–2412.
URL: http://www.hsph.harvard.edu/james-robins/files/2013/03/correcting-1994.pdf

Robins, James M. 1997. Causal Inference from Complex Longitudinal Data. In *Latent Variable Modeling and Applications to Causality*, ed. M. Berkane. Vol. 120 of *Lecture Notes in Statistics* New York: Springer-Verlag pp. 69–117.
URL: http://biosun1.harvard.edu/~robins/cicld-ucla.pdf

Robins, James M. 2003. Semantics of causal DAG models and the identification of direct and indirect effects. In *Highly Structured Stochastic Systems*, ed. P. J. Green, N. L. Hjort and S. Richardson. Oxford University Press pp. 70–81.

Robins, James M., Miguel A. Hernán and Babette A. Brumback. 2000. "Marginal Structural Models and Causal Inference in Epidemiology." *Epidemiology* 11(5):550–560.
URL: http://www.jstor.org/stable/3703997

Robins, James M. and Sander Greenland. 1992. "Identifiability and exchangeability for direct and indirect effects." *Epidemiology* 3(2):143–155.

Rosenbaum, Paul R. 1984. "The Consquences of Adjustment for a Concomitant Variable That Has Been Affected by the Treatment." *Journal of the Royal Statistical Society. Series A (General)* 147(5):656–666.

Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 6:688–701.

Rubin, Donald B. 2004. "Direct and Indirect Causal Effects via Potential Outcomes." *Scandinavian Journal of Statistics* 31(2):161–170.

VanderWeele, Tyler J. 2009. "Mediation and mechanism." *European Journal of Epidemiology* 24(5):217–224.

VanderWeele, Tyler J. 2011. "Controlled Direct and Mediated Effects: Definition, Identification and Bounds." *Scandinavian Journal of Statistics* 38(3):551–563.

VanderWeele, Tyler J. 2014. "A Unification of Mediation and Interaction: A 4-Way Decomposition." *Epidemiology* 25(5):749–761.

VanderWeele, Tyler J and Eric J Tchetgen Tchetgen. 2014. "Attributing Effects to Interactions." *Epidemiology* 25(5):711–722.

Vansteelandt, Sijn. 2009. "Estimating Direct Effects in Cohort and Case–Control Studies." *Epidemiology* 20(6):851–860.

Vansteelandt, Sijn. 2010. "Estimation of controlled direct effects on a dichotomous outcome using logistic structural direct effect models." *Biometrika* 97(4):921–934.

Vansteelandt, Sijn and Marshall Joffe. 2014. "Structural Nested Models and G-estimation: The Partially Realized Promise." *Statistical Science* 29(4):707–731.

Vansteelandt, Sijn and Tyler J VanderWeele. 2009. "Conceptual issues concerning mediation, interventions and composition." pp. 1–12.

A R code for Fearon and Laitin (2003)

In this appendix, we demonstrate how to implement the sequential g-estimation using the R statistical computing environment.

First, we load the Fearon and Laitin (2003) data, subset it as the authors did, and run the first stage regression. This regression includes both instability and any variable that is post-treatment to ethnic fractionalization.

```
fear <- read.dta("repdata.dta")
fear <- fear[which(fear$onset < 4),]

## first state (to get effect of instability)
first <- lm(onset ~ warl +  gdpenl + pan.lag(lpop, ccode) + lmtnest
  + ncontig + Oil + nwstate + instab + polity2l + ethfrac + relfrac,
  data = fear)
```

Then, in order to estimate the ACDE of ethnic fractionalization, we simply take the estimated coefficient on instability, multiply it by each unit's observed instability, and subtract that from the observed dependent variable. Here we do this in one regression. Note that this regression excludes any of the post-treatment variables from the first stage and only includes baseline variables.

```
## second stage (CDE of ethfrac net instab)
direct <- lm(I(onset - coef(first)["instab"]*instab) ~ lmtnest
  + ncontig  + Oil  + ethfrac + relfrac, data = fear)
```

While this regression will accurately estimate the point estimate for the ACDE under the above assumptions, but the standard errors from this regression will be incorrect. In particular, they will ignore the first stage completely. To get correct standard errors, we could bootstrap the entire process.

```
## bootstrap the SEs
boots <- 1000
fl.boots <- rep(NA, times = boots)

for (b in 1:boots) {
  fear.star <- fear[sample(1:nrow(fear), replace = TRUE),]
  boot.first <- lm(onset ~ warl +  gdpenl + pan.lag(lpop, ccode)
    + lmtnest + ncontig + Oil + nwstate + instab + polity2l + ethfrac
    + relfrac, data = fear.star)
  boot.direct  <- lm(I(onset - coef(boot.first)["instab"]*instab) ~ lmtnest
    + ncontig  + Oil  + ethfrac + relfrac, data = fear.star)
  fl.boots[b] <- coef(boot.direct)["ethfrac"]
}

sd(fl.boots)
```

B Completely mediated effects and the CDE

VanderWeele (2011) stated but did not formally prove that if M and W completely mediate the effect of A on Y, then any controlled direct effect of A on Y with $M = m$, $ACDE(m)$, must be due to an indirect effect of A on Y through W. First, let us define what complete mediation entails.

Definition 1 (Complete mediation). *A set of variables* $Z = \{Z_1, \ldots, Z_k\}$ completely mediates *the effect of A on Y if, for all values* $\{z_1, \ldots, z_k\} \in \mathcal{Z}$,

$$Y_i(a, z_1, \ldots, z_k) = Y_i(z_1, \ldots, z_k),$$

where $\mathcal{Z}$ *is the support of Z.*

Complete mediation is a common idea in the social sciences, where it is mostly seen in an instrumental variable design. There, the exclusion restriction assumes

that the effect of the instrument on the outcome is completely mediated by the treatment. Philosophically, it is unclear if all effects are possibly completely mediated by some set of variables or whether there are effects for which no completely mediating set exists. We do not answer this question. We assume that such a set exists and that we can partition it into two subsets: M the potential mediator we wish to test and W, the set of all other mediators. The goal of this analysis is to show that there is some effect that is not through M—that is, that there is some indirect effect through W. Because W is possibly multivariate with unobserved components, it may not be possible to determine what part of W mediates the effect. Thus, this approach represents a falsification test of sorts.

Proposition 1. *If the effect of A on Y is completely mediated by $Z = M, W$ and the consistency assumption holds for all potential outcomes, then the average controlled direct effect with M fixed is also an indirect effect of W:*

$$E[Y_i(a, m) - Y_i(a', m)] = E[Y_i(a, m, W_i(a, m)) - Y_i(a, m, W_i(a', m))].$$

Proof.

$$\begin{aligned} E[Y_i(a, m) - Y_i(a', m)] &= E[Y_i(a, m, W_i(a, m)) - Y_i(a', m, W_i(a', m))] \\ &= E[Y_i(a, m, W_i(a, m)) - Y_i(a, m, W_i(a', m)) \\ &\quad + Y_i(a, m, W_i(a', m)) - Y_i(a', m, W_i(a', m))] \\ &= E[Y_i(a, m, W_i(a, m)) - Y_i(a, m, W_i(a', m))] \end{aligned}$$

The first equality follows from the consistency assumption and the last equality follows from the definition of complete mediation. □

Note that this is a type of natural indirect effect—natural because the other mediators, W, are allowed to take their natural value under a, a', and m. Of course this indirect effect fixes the value of M to m, so that it ignores any potential interaction

between the indirect effect size and the natural value of M. It is also for this reason that this result holds whether M affects W, W affects M, or they are independent.

C Consistent Variance estimation for the ACDE in linear blip-down estimation

Let W_i be the $1 \times k$ vector of variables in the first stage of the blip-down estimation. In the paper above, we took this to include A_i, M_i, X_i, and Z_i, but here we will be more general so as to allow interactions and possible non-nesting of the first and second stages. Let V_i be the $1 \times p$ vector of variables in the second stage, direct effect model. Obviously, this includes A_i, but might also include baseline covariates (and interactions with baseline covariates) as well. Let $M_i \subset W_i$ be the vector of mediators, functions of mediators, and interactions between the mediators and the treatment or baseline covariates. This vector will be the vector of variables defined by the blip-down function for m. Let M_i have dimension k_m. We gather each of these row vectors in matrices W, V, and M, so that W for instance is an $n \times k$ matrix.

Let α be the vector of regression coefficients for the first model, α_m be the sub-vector of coefficients for M_i, and β be the vector of coefficients for the direct effect model. Given linear models for each stage, we can write the regression errors $u_{i1}(\alpha) = Y_i - W_i\alpha$ and $u_{i2}(\beta, \alpha) = Y_i - M_i\alpha_m - V_i\beta$. Let $\widehat{\alpha}$ be the estimator for α based on the sample moment conditions $n^{-1}\sum_i W_i^T u_{i1}(\widehat{\alpha}) = 0$ and $\widehat{\beta} = \widehat{\beta}(\widehat{\alpha})$ be the estimator based on the sample moment condition $n^{-1}\sum_i V_i^T u_{i2}(\widehat{\beta}, \widehat{\alpha}) = 0$. These are simply the OLS estimates from the first and second stages and $u_{i1}(\widehat{\alpha})$ and $u_{i2}(\widehat{\beta}, \widehat{\alpha})$ are the residuals.

Under standard theory (Newey and McFadden, 1994), Assumptions 1 and 2, and the assumption of correct linear models, we can show that $\widehat{\beta}$ is asymptotically Normal with asymptotic variance:

$$\text{Var}\left[\widehat{\beta}\right] = \left(E[V_i^T V_i]\right)^{-1} E\left[g_i g_i^T\right] \left(E[V_i^T V_i]\right)^{-1}, \tag{35}$$

with

$$g_i = V_i^T u_{i2} - F\left(E[W_i^T W_i]\right)^{-1} W_i^T u_{i1} \tag{36}$$

Here, $F = E[-V_i^T \widetilde{W}_i]$, where $\widetilde{W}_i = [M_i \;\; 0]$ is the vector of W_i with all non-M_i entries set to 0. To prove this, one only need note that the population moment conditions here are $E[V_i^T u_{i2}] = 0$ and $E[W_i^T u_{i1}] = 0$. Using the above assumptions and these moment conditions into Theorem 6.1 of Newey and McFadden (1994, p. 2178) yields (35).

To derive a consistent estimator for the variance, we simply plug sample versions of the population expectations in (35). Under regularity conditions $V^T V/n \xrightarrow{p} E[V_i^T V_i]$ and $W^T W/n \xrightarrow{p} E[W_i^T W_i]$ and we can use

$$\widehat{F} = -\frac{1}{n}\sum_i V_i^T \widetilde{W}_i, \tag{37}$$

which is consistent for F. Finally, we plug in the residuals to form:

$$\widehat{g}_i = V_i^T u_{i2}(\widehat{\beta}, \widehat{\alpha}) - \widehat{F}\left(W^T W\right)^{-1} W_i^T u_{i1}(\widehat{\alpha}). \tag{38}$$

Finally, we can combine each of these to form consistent variance estimator:

$$\widehat{\text{Var}}\left[\widehat{\beta}\right] = (V^T V)^{-1}\left(n^{-1}\sum_i \widehat{g}_i \widehat{g}_i^T\right)(V^T V)^{-1} \tag{39}$$

Note that this variance estimator is "robust" in the sense that it is consistent even if there is heteroskedasticity in either model. Given the structure of g_i and F, the variance of $\widehat{\beta}$ with α estimated will always be higher than if we were to have knowledge about the true α.

D Bias formulas and sensitivity analysis details

Let $\Delta(V_i|W_i) \equiv V_i - \widehat{E}[V_i|W_i]$ be the residuals of a regression of V_i on W_i By the Frisch-Waugh Theorem, we can write the estimated coefficient of M_i on Y_i as the following:

$$\widehat{\alpha}_2 = \alpha_2 + \frac{\sum_{i=1}^n \varepsilon_{iy} \Delta(M_i | Z_i, A_i, X_i)}{\sum_{i=1}^n \Delta(M_i | Z_i, A_i, X_i)}$$

Note that $\frac{1}{n}\sum_i \Delta(M_i|Z_i, A_i, X_i)^2$ converges in probability to $\text{Var}[\varepsilon_{im}]$ and $\frac{1}{n}\sum_i \varepsilon_{iy}\Delta(M_i|Z_i, A_i, X_i)$ converges to $\text{Cov}[\varepsilon_{iy}, \varepsilon_{im}]$. Combining these two facts with Slutsky's theorem gives the following:

$$\begin{aligned}
\text{plim}\ \widehat{\alpha}_2 &= \alpha_2 + \frac{\text{Cov}[\varepsilon_{iy}, \varepsilon_{im}]}{\text{Var}[\varepsilon_{im}]} && (40)\\
&= \alpha_2 + \frac{\rho\sigma_y\sigma_m}{\sigma_m^2} && (41)\\
&= \alpha_2 + \frac{\rho\sigma_y}{\sigma_m} && (42)
\end{aligned}$$

Let the true blipped-down outcome be $\tilde{Y}_i = Y_i - \alpha_2 M_i$. We can write $\tilde{Y}_i = \beta_0 + \beta_1 A_i + X_i^T\beta_2 + \eta_i$, where β_1 is the ACDE. Let $\widehat{\beta}_1$ be the coefficient from the regression of $\widehat{\tilde{Y}}_i$ on A_i and X_i. By the Frisch-Waugh theorem, we have:

$$\begin{aligned}
\widehat{\beta}_1 &= \frac{\sum_{i=1}^n (Y_i - \widehat{\alpha}_2 M_i)\Delta(A_i|X_i)}{\sum_{i=1}^n \Delta(A_i|X_i)^2} && (43)\\
&= \frac{\sum_{i=1}^n (\widetilde{Y}_i - \alpha_2 M_i + \widehat{\alpha}_2 M_i)\Delta(A_i|X_i)}{\sum_{i=1}^n \Delta(A_i|X_i)^2} && (44)\\
&= \frac{\sum_{i=1}^n \widetilde{Y}_i \Delta(A_i|X_i)}{\sum_{i=1}^n \Delta(A_i|X_i)^2} - \frac{(\alpha_2 - \widehat{\alpha}_2)\sum_{i=1}^n M_i \Delta(A_i|X_i)}{\sum_{i=1}^n \Delta(A_i|X_i)^2} && (45)
\end{aligned}$$

Let $M_i = \widetilde{\delta}_0 + \widetilde{\delta}_1 A_i + X_i^T\widetilde{\delta}_2 + \widetilde{\varepsilon}_{im}$ be the regression of M_i on A_i and X_i. Basic regression results establish that $\sum_i \widetilde{Y}_i\Delta(A_i|X_i)/\sum_i \Delta(A_i|X_i)^2$ converges to β_1 and $\sum_i M_i\Delta(A_i|X_i)/\sum_i \Delta(A_i|X_i)^2$ converges to $\widetilde{\delta}_1$. And given our above results, we have that $\alpha - \widehat{\alpha}_2$ converges to $\rho\sigma_y/\sigma_m$. Again using repeated applications of Slutsky's theorem, we can derive the asymptotic bias:

$$\text{plim}\ \widehat{\beta}_1 = \beta_1 - \frac{\rho\sigma_y\widetilde{\delta}_1}{\sigma_m}$$

Of course, σ_y is not identified due to the confounding. We take a similar approach to Imai, Keele and Yamamoto (2010) and note the following relationships between the various parameters:

$$\text{Var}[\widetilde{\varepsilon}_{iy}] = \widetilde{\sigma}_y^2 = \alpha_2^2\sigma_m^2 + \sigma_y^2 + 2\rho\alpha_2\sigma_m\sigma_y \tag{46}$$

$$\text{Cov}[\widetilde{\varepsilon}_{iy}, \varepsilon_{im}] = \widetilde{\rho}\widetilde{\sigma}_y\sigma_m = \alpha_2\sigma_m^2 + \rho\sigma_m\sigma_y \tag{47}$$

Solving for σ_y, we find that $\sigma_y = \widetilde{\sigma}_y\sqrt{(1-\widetilde{\rho}^2)/(1-\rho^2)}$. Plugging this into above yields the asymptotic bias formula.

To complete the proof note that (12) implies that $\widetilde{\delta}_1$ is identified from a regression of M_i on A_i and X_i. Under the LSEM and (12), we can use the

$$\widehat{\widetilde{\sigma}}_y = \sqrt{\text{Var}[\widehat{\widetilde{\varepsilon}}_{iy}]} = \sqrt{\sum_i (Y_i - \widehat{\widetilde{\alpha}}_0 - \widehat{\widetilde{\alpha}}_1 A_i - X_i^T\widehat{\widetilde{\alpha}}_3 - Z_i^T\widehat{\widetilde{\alpha}}_4)^2},$$

which is consistent for $\widetilde{\sigma}_y$. Furthermore,

$$\widehat{\sigma}_m = \sqrt{\text{Var}[\widehat{\varepsilon}_{im}]} = \sqrt{\sum_i (M_i - \widehat{\delta}_0 - \widehat{\delta}_1 A_i - X_i^T\widehat{\delta}_2 - Z_i^T\widehat{\delta}_3)^2}$$

which is consistent for σ_m. Finally, we can estimate $\widetilde{\rho}$ with the correlation between the residuals $\widehat{\widetilde{\varepsilon}}_{iy}$ and $\widehat{\varepsilon}_{im}$. Thus, given ρ the asymptotic bias of $\widehat{\beta}_1$ is identified and we can use this to identify β_1.

To get standard errors and confidence intervals for the sensitivity analysis, it is easier to correct for the bias in $\widehat{\alpha}_2$ and pass this bias-corrected estimate to the second stage. Then, the variance estimator of C consistently estimates the variance of $\widehat{\beta}$ as if the first stage were correctly specified. That is, it is the correct variance under the assumption that we have correctly chosen ρ.

Finally, note that we can reparameterize this sensitivity analysis to be as a function of the residual variation explained by unmeasured confounding. To see this, we introduce an unmeasured confounder, U_i:

$$\varepsilon_{iy} = \alpha_u U_i + \varepsilon_{iy}^* \tag{48}$$

$$\varepsilon_{im} = \delta_u U_i + \varepsilon_{im}^* \tag{49}$$

With these in hand, we can define the partial R^2 for U_i in terms of the outcome and the mediator:

$$R^2_y = 1 - \frac{\text{Var}[\varepsilon^*_{iy}]}{\text{Var}[\varepsilon_{iy}]} \tag{50}$$

$$R^2_m = 1 - \frac{\text{Var}[\varepsilon^*_{im}]}{\text{Var}[\varepsilon_{im}]} \tag{51}$$

These values represent the share of the unexplained variance in Y_i and M_i, respectively, that U_i explains. As shown by Imai, Keele and Yamamoto (2010), we have the following relationship between ρ and these partial R^2 values: $\rho^2 = R^2_y R^2_m$. Thus, we can vary these parameters, which imply differing values of ρ and consequently differing levels of bias. The advantage of this parameterization of the sensitivity analysis is that the partial R^2 may be more natural to interpret. For instance, we can compare them to the partial R^2 values of observed covariates in X_i and Z_i in order to gauge their relative magnitude (Imbens, 2004).

E Simulation setup

Here we present the simulation setup we discussed in Section 4 and present results from entire set of draws as opposed to only a single draw. This simulation is not meant to prove any property of any estimator—these results are largely known and have been established analytically. Instead, we show these for illustrative purposes.

$$\begin{aligned}
N &= 500 \\
A_i &\sim N(50, 15^2) \\
Z_i &\sim N(50, 15^2) \\
M_i &\sim N(0.5A_i + 0.5Z_i, 5^2) \\
Y_i &\sim N(75 - 0.5Z_i, 5^2)
\end{aligned}$$

We took 10,000 draws from this data generating process and ran three estimators on the samples. First, we ran a simple unconditional model of Y_i on A_i. Next, we ran a conditional model of Y_i on A_i and M_i. Finally, we applied the sequential

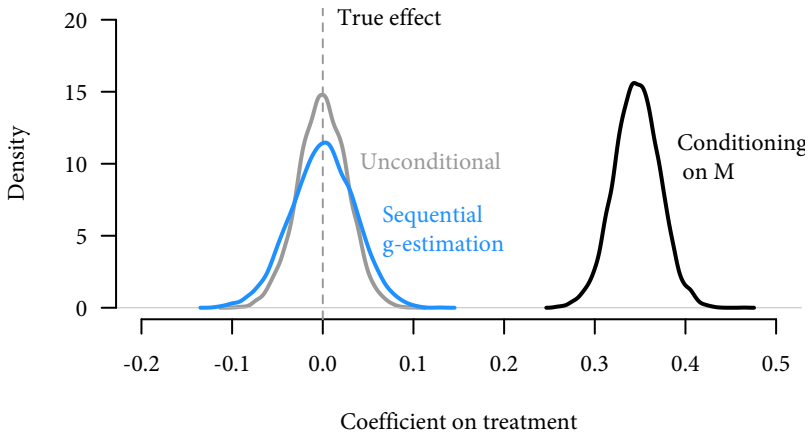


Figure 9: Simulated sampling distribution of the post-treatment bias estimator that conditions on M_i and the unbiased estimator that does not condition on M_i. Conditioning on a post-treatment covariate in this case produces serious bias.

g-estimation to estimate the direct effect of A_i, using Z_i as an intermediate confounder. We plot the results in Figure 9. Given the data generating process, it is unsurprising that both the unconditional and sequential g-estimation approach recover the truth on average, while conditioning on M_i induces very severe post-treatment bias. Note also that the sequential g-estimator has slightly higher variance than unconditional approach, which makes sense because the unconditional estimator is taking advantage of an addition restriction in this example: no effect of A_i on Z_i. If that were not true, then the unconditional estimator would also be biased.