How Can We Estimate the Effectiveness of Institutions? Solving the Post-Treatment versus Omitted Variable Bias Dilemma*

Michaël Aklin† Patrick Bayer‡

January 10, 2017

Abstract

Post-treatment confounding variables create a dilemma for applied researchers. Including these variables in regression models generates post-treatment bias. Excluding them generates omitted variable bias. Traditional estimation strategies can therefore not provide unbiased estimates in the presence of such variables. This has significant implications from a research perspective, because it also means that traditional regression methods offer biased estimates of the direct and indirect effects of an intervention, even though these are important to test competing theories. We suggest a simple estimation algorithm that can recover unbiased estimates of the direct, indirect, and total effect of an intervention when other estimation strategies fail to do so. We call it the total effect decomposition (TED) approach. This approach is particularly useful when the treatment variable has a diffuse effect, as is often the case when we study the effects of policies and institutions in international relations and comparative politics.

Keywords: post-treatment bias; omitted variable bias; direct effects; indirect effects; total effect decomposition.

*We are grateful to Pablo Barberá, Emine Deniz, Jude Hays, Marc Meredith, Matto Mildenberger, and David Steinberg for useful comments on earlier versions of this paper. We also thank audience members at the 16th Annual Policy Conference at the University of Pittsburgh and at ISA for suggestions. All errors are ours.

†Department of Political Science. University of Pittsburgh. 4814 WWPH, 230 S. Bouquet Street, Pittsburgh, PA 15260, USA. Email: aklin@pitt.edu.

‡School of Social and Political Sciences. University of Glasgow. Adam Smith Building, 40 Bute Gardens, Glasgow, G12 8RT, Scotland, UK. Email: patrick.bayer@glasgow.ac.uk

1 Introduction

As social scientists, we devote considerable efforts to estimate the effect of an intervention on a given outcome. Intervention, here, such as the adoption of a specific policy, refers to our independent variable of interest. Typically, we then define a statistical model and estimate the parameter of the intervention with the appropriate method. Among many threats, we are particularly worried about omitted variable bias. As a result, the search for sound identification strategies, whether by relying on experiments or quasi-experiments, has instilled valuable creativity in political science (Samii, 2016). For many reasons, however, not all interesting research hypotheses can be tested through these means and, more often then not, we need to rely on observational data. In this case, because the intervention has not been randomly assigned, control variables are included to our statistical model to ensure that the main estimate does not accidentally capture confounding factors.

This approach however leads to misleading findings. The first issue, which we leave aside in this paper, is that knowing the correct model is hard, despite a boom of techniques to reduce model dependence (Ho et al., 2007; Iacus, King, and Porro, 2012). The second, less well-known problem, is that even if we knew the true model, it may be impossible to recover unbiased estimates with traditional estimation methods because of the causal sequence of events. When a variable mediates our intervention, then the statistical model can neither omit the variable (since it is a confounder) nor can it include the variable (since it is post-treatment, that is, causally following our intervention). The former runs into problems of omitted variable bias, the latter leads to post-treatment bias. This is what Angrist and Pischke (2008, 64) call a "bad control."

The issue is ubiquitous in political science across subfields. Acharya, Blackwell, and Sen (2016, 512) find that "two-thirds of empirical papers in political science that make causal claims condition on posttreatment variables" and Montgomery, Nyhan, and Torres (2016) show that the problem is even relevant in randomized experiments. How to deal with post-treatment confounders remains, in King's (2010) words, one of the "hard questions" in the social sciences.

This, in turn, sheds light on an even bigger issue: the need to distinguish between the *direct*, *indirect*, and *total effect* of an intervention (Imai et al., 2011; Acharya, Blackwell, and Sen, 2016). The direct effect of an intervention, or DEI, refers to the unmediated consequences of an intervention.

These are often the estimates that are reported in empirical studies and are typically associated with the regression coefficient of the intervention indicator. When a government introduces a minimum wage policy, for example, the direct effect captures the direct increase in the country's income levels. But an intervention typically also operates through a (post-treatment) mediator, over and above the direct effect (Gerber and Green, 2012; Imai et al., 2011). We call this the *indirect* effect of the intervention, or IEI. The *total* effect of an intervention (TEI), then, is the sum of the DEI and IEI. Distinguishing between these quantities is desirable because it helps us assess the validity of competing theories by decomposing an intervention's total effect into its constituent parts. In the above example, a minimum wage policy not only increases income levels directly through cash transfers, but also indirectly through changing incentives of labor market participation.

To summarize, our problem is simple. We would like to get reliable estimates of the direct, indirect, and total effects of an intervention. Unfortunately, finding a solution is difficult. Even if our theory offers clear guidance as to what the (post-treatment confounding) mediator is, since it is measured post-treatment, we cannot get unbiased estimates of *all* these quantities. We are trapped between a rock (post-treatment bias) and a hard place (omitted variable bias).

In what follows, we contribute to the literature in two ways. First, we spell out exactly how post-treatment bias can affect estimates. We focus on two particularly important assumptions that affect our empirical analysis: sequential ignorability and exclusion restriction. Then, we offer a review of various existing approaches to deal with post-treatment bias: an instrumental variable (IV) approach and a recently advanced approach to estimate the average controlled direct effect (ACDE) (Acharya, Blackwell, and Sen, 2016). We clarify the assumptions that need to be met for these to offer unbiased estimates. We also note that these approaches can only estimate direct effects and, by construction, fail to help us retrieve the indirect and total effect of an intervention.

Second, we suggest a new estimation algorithm that can recover the direct, indirect, and total effect. We call this the *total effect decomposition* (TED) approach. In the spirit of Abadie, Diamond, and Hainmueller's (2010) synthetic control method, our total effect decomposition estimates rely on untreated, pre-intervention observations to reconstruct the unobserved counterfactual among treated observations. Based on these counterfactuals, we can generate unbiased estimates of the

intervention effect even in the presence of post-treatment confounders. We spell out the conditions under which these different estimators perform best. Finally, we apply our TED approach to examine the problem of the effect of democratization on FDI. Jensen (2003) noted that the effect of establishing democratic institution could be diffuse. For one, it could be operate directly by increasing audience costs against illegal appropriation. But it could also be indirect: democracies implement better policies to attract foreign investments; for instance, they tend to implement laxer capital controls. We show that the effect of democratization operates mostly through audience costs and not through policies.

In practice, the value of the algorithm we propose in this paper critically depends on the quality of our theories. To tease out the direct, indirect, and total effect of an intervention, we need to have a good understanding of the data generating process, which only well specified theory can provide us with. From a normative perspective, this is a desirable feature, as any empirical technique will only be as powerful as the social science theories that underlie it. We thus join recent voices arguing in favor of better ties between theory and empirical work (Samii, 2016), and hope that our method will help applied researchers test new predictions from their work.

2 Defining the Problem

To fix ideas, let us begin by characterizing the data generating process (DGP) and the general setup we are operating in. We will also define our quantities of interest, that is, the direct, indirect, and total effect of an intervention.

2.1 Setup

Figure 1 presents the general form of the issue that we face. This directed acyclic graph relates our main independent variable of interest, D, an indicator variable of policy adoption, say, to our outcome variable of interest, Y. Part of the total effect of D on Y is direct and part is mediated through X. To make the model more relevant to applied researchers, we also include a traditional confounder, Z, and allow for another variable W, which can serve as an instrument for X assuming the conventional exclusion restriction holds.

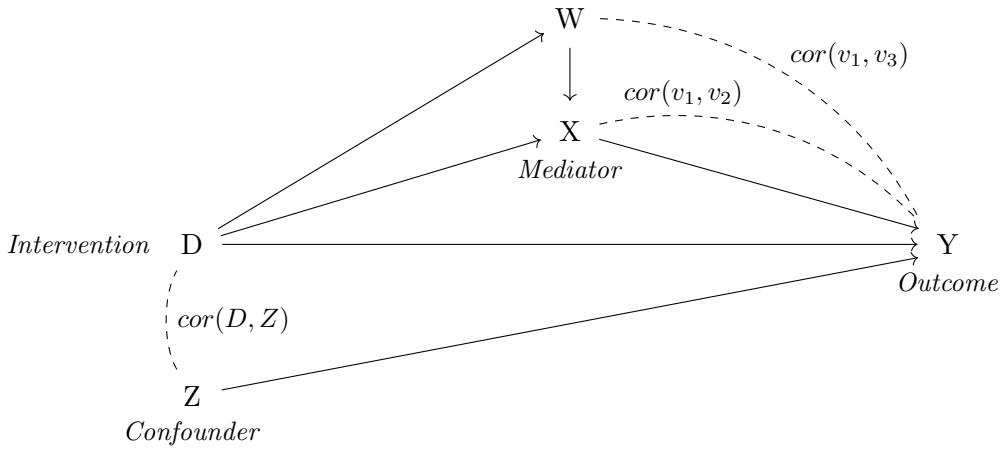


Figure 1: Directed acyclic graph showing our data generating process. Y is the outcome variable, D is our intervention of interest, and X is a mediator of the effect of D. Z denotes a classical confounder, and W serves as a possible instrument for X. Arrows indicate causal relationships and dotted lines denote variable ($cor(D, Z)$) and error correlations ($cor(v_1, v_2)$ and $cor(v_1, v_3)$).

We impose the following restrictions on this data generating process:

- $Y \in \mathbb{R}$ is the *outcome* (dependent) variable.
- $D \in \{0, 1\}$ is a binary *intervention* variable. "Zero" means the absence of the intervention, and "one" denotes the intervention is present in a given unit. Importantly, the intervention is *not* randomly assigned.
- $X \in \mathbb{R}$ is a post-treatment confounding variable, which we refer to as the *mediator*. Post-treatment means that in post-intervention periods only part—and not all—of X's variation is caused by variation in D. If this unmodeled variation, as captured by $cor(v_1, v_2)$ in Figure 1, is systematically correlated with the outcome Y, sequential ignorability does not hold.[1]
- Z is a confounding variable which we assume to be correlated with D, as otherwise including it into our model would be unnecessary. Indeed, without loss of generality, we assume $cor(D, Z) = 0.3$ throughout.

[1] In our case, sequential ignorability means: (i) D and Y are statistically independent conditional on Z, which is essentially an assumption about correct model specification; (ii) X and Y are statistically independent conditional on D, W, and Z. It is this second part of the sequential ignorability assumption which is violated for $cor(v_1, v_2) \neq 0$. See Acharya, Blackwell, and Sen (2016) and Imai et al. (2011) for details.

- W serves as a possible instrument for X. By construction, W affects X, so the validity of the instrument depends on the extent to which the exclusion restriction is met. W is a valid instrument if error terms v_1 and v_3 are uncorrelated.

Given this setup, how can we recover the parameters of this model? Specifically, how can we estimate the effect of D on Y? Of course, we could simply regress Y on D (and Z and W) with least squares to avoid including the post-treatment confounder X into our model. The problem with this approach is that ignoring X generates a biased estimate of the true effect of D. This is a classical case of omitted variable bias (Greene, 2008).

Could we then instead simply add X to our model and regress Y on D and X (and Z and W)? Unfortunately not. Because X is a post-treatment variable, including it creates a different kind of bias, typically referred to as post-treatment bias (Rosenbaum, 1984; King, 2010; Gerber and Green, 2012).

This is the dilemma between post-treatment and omitted variable bias, which is at the heart of this paper. We further distinguish between two sources where post-treatment bias can be coming from. Once we include the post-treatment confounder X and estimate the full model as shown in Figure 1, bias can come from an unmodeled relationship of X and Y. In post-intervention periods, D may no longer explain all the variation in X, leaving parts of it unmodeled. This results in non-zero error correlation $cor(v_1, v_2)$, biasing regression estimates away from their true values. We refer to this as type v_1v_2 post-treatment bias.

The second source of bias is due to an unmodeled relationship between W and Y. If W is correlated in a systematic way with the outcome variable, which is not captured by the statistical model, $cor(v_1, v_3)$ is non-zero and bias occurs. As we shall see below, assumptions about these error correlations determine which estimation strategy is the most suitable to get around the post-treatment and omitted variable bias dilemma. In keeping with the above, we call bias that come from this second source as v_1v_3 post-treatment bias.

2.2 Quantities of Interest

Before discussing possible solutions to the problem, let us first define our main quantities of interest which we seek to estimate.

We can rewrite our data generating process from Figure 1 with the following set of equations, where we include v_1, v_2, and v_3 as error terms:

$$Y = \alpha_1 + \beta_1 D + \gamma_1 X + \delta_1 Z + v_1 \tag{1}$$

$$X = \alpha_2 + \beta_2 D + \gamma_2 W + v_2 \tag{2}$$

$$W = \alpha_3 + \beta_3 D + v_3. \tag{3}$$

The direct effect of D is β_1. But as noted earlier, the direct effect is not the same as the total effect. Indeed, D affects Y through the mediator X. In addition, D also affects W, which in turn affects X. What, then, is the total effect of the intervention? To find this out, we substitute W from equation (3) into equation (2) and then again substitute this into equation (1), which—after rearranging—gives:

$$Y = [\alpha_1 + \gamma_1\alpha_2 + \gamma_1\gamma_2\alpha_3] + D[\beta_1 + \gamma_1\beta_2 + \gamma_1\gamma_2\beta_3] + \delta_1 Z + [v_1 + \gamma_1 v_2 + \gamma_1\gamma_2 v_3]$$

We can then compute the *total effect* of the intervention (TEI) as

$$\text{TEI} \equiv \frac{\partial Y}{\partial D} = \beta_1 + \gamma_1\beta_2 + \gamma_1\gamma_2\beta_3,$$

which can be decomposed into the *direct effect* of the intervention (DEI) and the *indirect effect* of the intervention (IEI):

$$\text{TEI} \equiv \text{DEI} + \text{IEI}, \quad \text{where DEI} \equiv \beta_1 \text{ and IEI} \equiv \gamma_1\beta_2 + \gamma_1\gamma_2\beta_3.$$

To avoid trivial cases, we assume that D has both a direct and indirect effect on Y.[2] Notably,

[2] We require $\beta_1 \neq 0$ for the direct effect. For the indirect effect, $\gamma_1 \neq 0$ and $\beta_2/\beta_3 \neq -\gamma_2$ have to hold.

this does not preclude the total effect to be zero if DEI $= -$IEI.

3 Estimation Strategies

How can we now recover unbiased estimates of the direct, indirect, and total effects? As we stated above, the naive approach of regressing Y on D (with or without X) immediately leads to biased estimates. We leave these results to the appendix (Section S1). Here, we focus on three more sophisticated estimation strategies to deal with the issue. We briefly introduce each approach and also assess the assumptions needed to obtain unbiased DEI estimates, while discussing our most preferred estimation strategy—the Direct Effect Decomposition (TED) approach—in the next section.

3.1 Three Solutions to Post-Treatment Bias

We consider the following three estimation strategies: Instrumental variable (IV) approach, Average Controlled Direct Effect (ACDE) approach (Acharya, Blackwell, and Sen, 2016), and our new Total Effect Decomposition (TED) approach. These estimation strategies can produce unbiased estimates of the direct effect if the assumptions specified in Table 1 hold. Importantly, only the TED approach, which we introduce here, allows us to obtain estimates of the indirect and total effect.

The IV approach relies on the following intuition. We can get around post-treatment bias of the v_1v_2-type by avoiding to use the post-treatment confounding variable X as a regressor altogether. Instead of regressing Y on X (and confounder Z), we instrument X with W. If W is a valid instrument, the direct effect can be estimated without bias. In a standard 2-stage least squares (2SLS) IV model, the instrument's coefficient captures the direct effect of the intervention. For this coefficient to be unbiased, we simply need the error correlation of v_1 and v_3 to be zero, as per the standard exclusion restriction assumption. Without adding X as an independent variable to the second stage of the IV regression, any form of post-treatment bias that would come in through including X is inconsequential for DEI estimates. The IV approach thus breaks the correlation of the moderator X and the outcome variable Y (which induces v_1v_2-type post-treatment bias) by instrumenting X with W.

Specification	Estimand	Assumption
Instrumental variable (IV) approach		
(1) Regress $X = \alpha + \gamma W + v$ (2) Predict $\hat{X} = \hat{\alpha} + \hat{\gamma} W$ (3) Regress $Y = \alpha_1 + \beta_1 D + \gamma_1 \hat{X} + \delta_1 Z + v_1$	DEI ($= \hat{\beta}_1$)	Exclusion restriction $\text{Cor}(v_1, v_3) = 0$
Average controlled direct effect (ACDE) approach		
(1) Regress $Y = \alpha + \beta D + \gamma X + \delta Z + \epsilon W + v$ (2) Predict $\hat{Y} = \hat{\gamma} X$ (3) Demediate $\tilde{Y} = Y - \hat{Y}$ (4) Regress $\tilde{Y} = \alpha_1 + \beta_1 D + \delta_1 Z + \epsilon_1 W + v_1$	DEI ($= \hat{\beta}_1$)	Sequential ignorability $\text{Cor}(v_1, v_2) = 0$
Total effect decomposition (TED) approach		
(1) Regress $Y_{D=0} = \alpha + \delta Z_{D=0} + v$ (2) Predict $\hat{Y}_{D=1} = \hat{\alpha} + \hat{\delta} Z_{D=1}$ (3) Calculate $\widehat{\text{TEI}} = E[\hat{Y}_{D=1} - Y_{D=1}]$	TEI	None
(1) Regress $Y_{D=0} = \alpha + \gamma X_{D=0} + \delta Z_{D=0} + v$ (2) Predict $\hat{Y}_{D=1} = \hat{\alpha} + \hat{\gamma} X_{D=1} + \hat{\delta} Z_{D=1}$ (3) Calculate $\widehat{\text{DEI}} = E[\hat{Y}_{D=1} - Y_{D=1}]$	DEI	Exclusion restriction for pre-intervention periods $\text{Cor}(v_1, v_3)_{D=0} = 0$
(1) Calculate $\widehat{\text{IEI}} = \widehat{\text{TEI}} - \widehat{\text{DEI}}$	IEI	Same as DEI

Table 1: Estimation strategies to address post-treatment bias. The true, data generating model is the one in Figure 1 and equations (1)-(3). Column (3) specifies the assumption about error correlations that must be meet, so that the quantities of interest in column (2) can be estimated without bias. *Notes:* DEI=direct effect; IEI=indirect effect; TEI=total effect. Variables and coefficients that are superscripted with " ˆ " denote predicted values and estimated coefficients, respectively. We suppress subscripts in auxiliary regressions to indicate that these coefficients are different from the ones in the data generating process in equations (1)-(3). As discussed in the text, $D = 0$ and $D = 1$ denote pre-intervention and post-intervention periods.

Another way to derive unbiased estimates of the direct effect in the presence of post-treatment bias has recently been introduced by Acharya, Blackwell, and Sen (2016). Their ACDE estimator takes a different tack in that it essentially demediates the dependent variable to "net out" the confounding effect of the mediator in post-intervention periods. This then allows to estimate the direct effect without the need to control for post-intervention variables. By not having to include W (or X) in the final estimation of the demediated outcome $\tilde{Y}$, we can therefore avoid $v_1 v_3$-type bias in DEI estimates. The ACDE approach is particularly useful when (as is often the case) good

instruments are hard to come by with as it is insensitive to correlation of W and Y.

For ACDE to work though, sequential ignorability needs to hold. This assumption consists of two parts. First, our treatment and outcome variables (after controlling for confounder Z) need to be statistically independent, which is to say that our model is correctly specified. As indicated above, we take this for granted. Second, sequential ignorability requires that conditional on D, Z, and W, our mediator and outcome variables are uncorrelated, which translates to $cor(v_1, v_2) = 0$ in our context.

Finally, the Total Effect Decomposition (TED) algorithm allows us to recover unbiased estimates of *all* three main quantities of interest if a somewhat relaxed version of the exclusion restriction holds: We only need the error correlation of v_1 and v_3 to be zero in *pre*-intervention periods for TED to work. To foreshadow, the TED approach dodges problems of post-treatment bias by not relying on any post-intervention variables at all. Instead, TED only uses (bias-free) pre-intervention period observations to predict a post-intervention counterfactual, which after first differencing observed and predicted values, provides unbiased estimates of the direct, indirect, and total effect.[3]

In summary, all three approaches are able to produce unbiased estimates of the direct effect. However, assumptions about error correlations differ, as summarized in Table 1. The IV approach needs the standard exclusion restriction to hold, while ACDE obtains unbiased direct effects under sequential ignorability. Finally, the TED approach relaxes the exclusion restriction because it only requires zero error correlation of v_1 and v_3 in pre-intervention periods. None of these assumptions are testable with data. The usefulness of each approach hence depends on a researcher's belief about whether it seems more that X and Y *or* W and Y are uncorrelated. Clearly though, TED excels in that it is the *only* approach that can estimate the direct, indirect, and total effect without bias.

3.2 Monte Carlo Simulations for IV and ACDE Approaches

Before introducing the TED approach, we assess the bias in DEI estimates from the IV and ACDE estimation strategies when the above assumptions are not met. We do so by running Monte Carlo

[3]Relative to Acharya, Blackwell, and Sen's (2016) ACDE approach, we are able to estimate the indirect effect, as we do not treat the mediator's effect on Y as a "nuisance" that is netted out through demediation.

simulations with randomly generated data for different levels of error correlations v_1, v_2 and v_1, v_3. Implementing the data generating process as shown in Figure 1 allows us to control the exact source of post-treatment bias independently.

For each combination of error correlations, we produce 1,000 time-series, cross-sectional data sets with $i = 100$ units over $t = 20$ time periods, for a total of $n = 2,000$ observations per data set. The *binary* intervention indicator changes from $D = 0$ to $D = 1$ for *all* observations in periods $t > 10$, perfectly splitting the data into ten pre-intervention and ten post-intervention periods right in the middle of the time-series.[4]

Plotting the distribution of direct effect estimates for the IV approach, Figure 2(a) shows that the unbiased DEI estimate of DEI $= \hat{\beta}_1 = 1$, indicated by the dashed line, can be correctly retrieved when the exclusion restriction assumption holds. Importantly, this result is unaffected by changes in the error correlation of v_1 and v_2. As argued above, as long as W serves as a valid instrument for X, correlation between the mediator and the outcome is inconsequential for the IV approach to produce unbiased estimates. In the bottom panel, we demonstrate that the percentage error in the direct effect estimate increases quickly (and in a linear fashion) for assumption violations. The 3-dimensional heatmap in Figure 2(b) suggests that at the ends of our $[-0.5, 0.5]$-interval, coefficient estimates can be biased by up to 200%.[5] This is, for example, compatible with (biased) coefficient estimates of the direct effect to be $+2$ (when $cor(v_1, v_3 = -0.5)$) or -1 (when $cor(v_1, v_2) = 0.5$), as illustrated in the top panel of Figure 2.

Next, for the average controlled direct effect (ACDE) approach, our simulations in Figure 3(a) attest that the direct effect can be estimated without bias, as Acharya, Blackwell, and Sen (2016) show, if sequential ignorability holds. Netting out the effect of the mediator on the dependent variable works as long as X and Y are uncorrelated. Loosely speaking, this estimation strategy can work as all indirect effects that are caused by changes in D are "absorbed" once the outcome variable Y is demediated. Indirect effects are essentially treated as a "nuisance." Again, the correlation of W and Y, as given by $cor(v_1, v_3)$, does not matter for the ACDE approach to produce unbiased

[4] The full details of the simulation setup are in Appendix S2.

[5] The choice of the $[-0.5, 0.5]$-interval for correlation coefficients is innocuous for our results, but convenient as it ensures for our variance-covariance matrix to be strictly positive definite. Also see Appendix S2.

estimates. Our heatmap in the bottom panel of Figure 3(b) shows that bias is however sizeable if sequential ignorability is violated. We find bias of up to 200%, or coefficient estimates that over- or underestimate the true effect by a factor of 2.

In summary, our Monte Carlo simulations offer some clear advice: The two most promising estimation strategies, the IV and ACDE approaches, can produce unbiased DEI estimates. However, estimates are heavily biased once the exclusion restriction (IV approach) or sequential ignorability (ACDE approach) assumptions fail to hold. As both assumptions are inherently non-testable, this is not a qualification we should take lightly. More limiting even, none of these two approaches can be used to recover estimates of the direct, indirect, and total effects of a (policy) intervention.

4 Total Effect Decomposition (TED) Approach

We now introduce our simple, yet powerful *total effect decomposition* (TED) approach as a solution to the above described problem of post-treatment bias. After building some basic intuition for why this approach works, we show how easy its implementation is and discuss key assumptions, using the same Monte Carlo simulations as above to illustrate our approach.

4.1 Intuition

The intuition behind the TED approach is compellingly straightforward: It takes the problem of post-treatment bias head-on by avoiding to estimate coefficients with post-intervention data altogether. Instead, we only use *pre-intervention* data to derive unbiased coefficient estimates, which we then use to estimate predicted values of Y. These predicted values serve as an estimate of the *counterfactual* outcomes, and first differencing predicted and observed values in post-intervention periods allows us to derive our desired quantities of interest.

Whether the TED approach produces an estimate of the DEI or TEI depends on the set of variables that are included when regressing Y on the pre-intervention data in the first step of the TED approach. Regressing Y on the confounder Z alone (this generalizes to a set of confounding variables) gives an estimate of the *total* effect. This makes intuitively sense as all the variation in Y must either come from variation in Z or variation caused by changes in the intervention D either

through direct (direct arrow of D to Y in Figure 1) or indirect channels (indirect path of D to Y through first W and X in Figure 1). Alternatively, regressing Y on Z *and* X produces an estimate of the direct effect because including X as an additional regressor "absorbs" all the variation in Y that comes through the indirect path of D to Y through W and X.[6] Controlling for X in our TED approach has the same effect as demediating the dependent variable in Acharya, Blackwell, and Sen's (2016) ACDE approach—in both cases, the indirect effects are treated as a nuisance we need to control for or net out.

We can finally derive an estimate of the indirect effect by simply calculating the difference of the total and direct effect estimates. The estimate of the IEI is unbiased when both the TEI and DEI are unbiased; we discuss the assumptions that guarantee unbiasedness further below.

4.2 Implementation of TED

A major strength of our approach is the ease with which it can be implemented in a standard regression framework. As introduced above, we index a variable's pre-intervention values with $D = 0$ and its post-intervention values with $D = 1$. So if an intervention occurred in period $t = 11$ of a time-series with 20 observations, as we assume in our Monte Carlo simulation setup, we would conveniently write $D = 0$ to denote the first ten values of a given variable; $D = 1$ would refer to the last ten observations in periods $t = 11, ..., 20$. This highlights the importance for our indicator variable to be binary, such as when it measures policy adoption or implementation, change in regime type from autocratic to democratic institutions, or a government's decision to join an international organization.

Total effect of intervention (TEI)

To estimate the TEI of an intervention, simply follow this 3-step procedure:

(1) Regress $Y_{D=0}$ on the confounding variable $Z_{D=0}$ for pre-intervention periods.

(2) Predict $\hat{Y}_{D=1}$ for post-intervention periods using post-intervention data $Z_{D=1}$ and the estimated intercept and slope coefficients.

[6] See Section S3 for a formal proof.

(3) Calculate the first difference of predicted (counterfactual) outcomes $\hat{Y}_{D=1}$ and observed values $Y_{D=1}$; averaging over this first difference produces an estimate of the TEI.

Direct effect of intervention (DEI)

The direct effect can be estimated using an almost identical procedure, while also controlling for the moderator X in the first setup of the above described algorithm.

(1) Regress $Y_{D=0}$ on the confounding variable $Z_{D=0}$ *and* the moderator $X_{D=0}$ for pre-intervention periods.

(2) Predict $\hat{Y}_{D=1}$ for post-intervention periods using post-intervention data $Z_{D=1}$ and $X_{D=1}$ and the estimated intercept and slope coefficients.

(3) Calculate the first difference of predicted (counterfactual) outcomes $\hat{Y}_{D=1}$ and observed values $Y_{D=1}$; averaging over this first difference produces an estimate of the DEI.

Indirect effect of intervention (IEI)

As the total effect is additive in its direct and indirect effects, an estimate of the indirect effect can simply be calculated as the difference of the TEI and DEI estimates. Our proposed approach essentially allows us to decompose the total effect into its constituent parts, and hence the name.

The procedure of the TED approach is summarized in Table 1, using the same notation as for the other two approaches to facilitate easy comparison.

4.3 Monte Carlo Simulations for TED Approach

Putting our approach to a first test, we replicate the same Monte Carlo simulations described above, with the major difference that the TED approach can produce estimates of all three quantities of interest. Figure 4(a) indeed plots the kernel densities of both the direct (blue) and total effect (gray) estimates. We do not explicitly plot the distribution of the indirect effects, which is simply the difference of two distributions that are shown. The expected true value for the DEI is 1 and for the TEI is 9.

As we can see, the estimate for the total effect is never biased. This is a remarkable feature of our TED approach. The direct effect is unbiased as long as a relaxed exclusion restriction assumption holds. Compared to the IV approach which requires the error correlation of v_1 and v_2 to be zero for *all* periods, TED only needs $cor(v_1, v_2) = 0$ in pre-intervention periods. As we are not using post-intervention observations to estimate coefficients in the first step of our algorithm, bias that occurs after the intervention kicks in is inconsequential for our estimates. Put differently, if we have reason to believe that X and Y are only correlated in post-intervention periods, our TED approach "automatically" always produces unbiased direct effect estimates, as it already does for the total effect.

Comparing the size of bias in DEI estimates, TED produces bias of up to 150% relative to the true value, as shown in Figure 4(b). While this is by no means small, the bias is still about a quarter smaller than the bias of both the IV and ACDE approaches, for comparable combinations of error correlations.

The TED approach is hence not just the only estimation strategy that can produce estimates of all three quantities of interest, that is, the total, direct, and indirect effect, but also requires a more "modest" assumption to ensure unbiased coefficients (at least relative to the IV approach). If the assumptions for unbiased estimates are not met, the size of the bias on the $[-0.5, 0.5]$-interval is also non-trivially smaller.

5 Application

We next demonstrate a concrete application of our method. Foreign direct investments (FDI) are believed to be a key driver of economic growth (de Mello, 1997; de Mello, 1999; Chowdhury and Mavrotas, 2006). FDI refers to cross-national private investments generally assumed to follow a long time horizon. These investments increase the amount of capital in a given country, which generates growth and employment. Given these attractive features, governments often seek to attract FDI. In his seminal work, Jensen (2003) argues that FDI is particularly likely to be targeted toward democracies. This is because democracies reduce political risks faced by international corporations. Once an investment is made, governments may be tempted to expropriate firms—a classical time

inconsistency problem (Kobrin, 1987). Jensen (2003, 594) offers two reasons why democracies are less likely to do so. First, democracies have more veto players, reducing the ability of the government of expropriating firms. Second, democratic governments could suffer from audience cost if they were deemed unreliable by voters.[7]

Empirically, Jensen (2003, 605) finds strong empirical support for his main claim: democracies receive more FDI. We focus on his cross-section time series results (Table 4), especially Model 10 and 11. The regression equation for the latter is:

$$\text{FDI (\% of GDP)}_{i,t} = \alpha_i + \tau_t + \beta\text{Democracy}_{i,t} + \gamma\text{Capital Controls}_{i,t} + \psi'\mathbf{W}_{i,t} + \theta'\mathbf{Z}_{i,t} + \varepsilon_{i,t}$$

where α are country fixed effects and τ are decade dummies. We discuss the content of the two vectors $\mathbf{W}$ and $\mathbf{Z}$ below.

Drawing on Polity data, an increase of the democracy index by one standard deviation (7.78) increases FDI per GDP by 0.16 percentage points. We are able to replicate perfectly these results using his replication package (Table 2).[8] Column (1) and (3) replicate Jensen's Table 4, Models 10 and 11, respectively. Since our approach does not work well with continuous interventions, we also verify that the results are robust when using the democracy indicator from Alvarez et al. (1996), which we label *Democracy (ACLP)*. Again, Jensen's results are very robust (columns 2 and 4). Switching from autocracy to democracy increases FDI by 0.38 percentage points.

In our view, one weakness in this analysis is the crucial role played by capital controls. Capital controls refer to any policy that facilitates or prevents capital flows across countries (King and Levine, 1993). The free movement of capital across borders may strengthen the willingness of investors to send their capital abroad, since it should mean that they will not be penalized if they ever want to repatriate their investments. Since capital controls are a policy, they are under the (constrained) control of policymakers. The problem is that a vast literature posits that democracies are more likely to implement liberal economic policies (Milner and Kubota, 2005; Frye and

[7] At the same time, Jensen acknowledges that autocracies may have advantages too, though he believes these to be overstated. Autocracies can offer better deals to firms, a cheaper labor force, a weaker regulations.

[8] We made one modification to Jensen's original dataset: the capital control variable he used is not publicly available. Instead, we use Karcher and Steinberg (2013).

Replication of Jensen (2003), Table 4

	Model 10	Model 10 (alt.)	Model 11	Model 11 (alt.)
	(1)	(2)	(3)	(4)
FDI (t-1)	0.364***	0.309***	0.321***	0.271***
	(0.069)	(0.066)	(0.066)	(0.063)
Market Size	−0.554	−0.614	−0.193	−0.103
	(0.462)	(0.453)	(0.466)	(0.458)
Development	0.834*	0.947**	0.492	0.376
	(0.466)	(0.445)	(0.503)	(0.487)
Growth	0.024***	0.023***	0.021**	0.018**
	(0.008)	(0.008)	(0.008)	(0.008)
Trade	0.006	0.007	0.007	0.009*
	(0.004)	(0.005)	(0.005)	(0.005)
Budget Deficit	−0.023**	−0.024**	−0.018*	−0.020*
	(0.011)	(0.011)	(0.011)	(0.011)
Gov Consumption	−0.039**	−0.045**	−0.031*	−0.036**
	(0.017)	(0.018)	(0.017)	(0.018)
Democracy (Polity)	0.021**		0.021**	
	(0.008)		(0.009)	
Democracy (ACLP)		0.380***		0.384***
		(0.101)		(0.098)
Capital Openness (K&S)			0.167***	0.208***
			(0.043)	(0.045)
Observations	1630	1584	1532	1522
Country Fixed Effects	Yes	Yes	Yes	Yes
Time Fixed Effects	Yes	Yes	Yes	Yes

Note: Panel corrected standard errors in parentheses. *p<0.1; **p<0.05; ***p<0.01

Table 2: Replication of Jensen (2003, Table 4). Unit of analysis is country-year.

Mansfield, 2004), and this is also true of capital controls (Brooks and Kurtz, 2007). It is therefore highly likely that capital openness is a post-treatment variable. Therefore, the results of Table 2, columns (1) and (3) are likely to suffer from omitted variable bias (capital controls have an effect on FDI), while columns (2) and (4) are likely to suffer from post-treatment bias.

The question then is: Do democracies receive more FDI because (a) they are more credible (audience cost), or (b) because they adopt more liberal policies (capital controls)? Audience cost is notoriously difficult to measure (Chaudoin, 2014). However, we can use our measurement strategy to at least assess the strength and relevance of capital controls.

We adapt Jensen's (2003) model to our framework (Figure 5). In our previous notation, democracy is the intervention D and capital controls are the mediating variable X. What, then, are $\mathbf{Z}$ and $\mathbf{W}$? Variables that are causally related to capital controls should be included in the vector $\mathbf{W}$ and those that are unrelated should be attached to $\mathbf{Z}$. We draw on the literature on the determinants

Estimates of the Direct, Indirect, and Total Effects

Estimand	Estimate	95% Confidence Interval
Direct effect of democracy ($\widehat{\text{DEI}}$)	0.45	[0.29,0.62]
Indirect effect of democracy ($\widehat{\text{IEI}}$) (through capital controls)	0.03	[-0.22,0.25]
Total effect of democracy ($\widehat{\text{TEI}}$)	0.48	[0.29,0.65]

Table 3: Estimates of the direct (DEI), indirect (IEI), and total (TEI) effect of democracy on FDI inflows (as a percentage of GDP). The mediating variable is an index of capital controls (Karcher and Steinberg, 2013). The set of variables in the vector **W** and **Z** are described in the text.

of capital controls to partition the confounding variables in either category. Specifically, we rely on Brooks and Kurtz (2007, 710), who analyze the economic and political sources of capital controls. We cross-check their explanatory variables and the ones used by Jensen (2003). Those present in both papers are incorporated in **W** while those that are not are included in the **Z** vector. That way we remain as close as possible to Jensen's original specification, which makes the results more comparable.[9] This enables us to define the two vectors as follows:

- **W**: development level (i.e., GDP per capita), growth, trade;
- **Z**: market size, budget deficit, government consumption, lagged FDI, and fixed effects.

We are then in a position to estimate all quantities of interest: DEI, IEI, and TEI. We follow the procedure described in Section 4. In order to construct confidence intervals for the total, direct, and indirect effects, we use non-parametric bootstrapping.[10] The results are reported in Table 3.

We find that democracy increases FDI when we consider the direct effect only (net of capital controls). The estimate of DEI is 0.45. The indirect effect, which is the effect of democracy through capital controls, is positive (0.03), but it is very small in comparison to the total effect. To further contrast the difference between the DEI and IEI, we plot the point estimates in Figure 6.

[9] In the appendix, we add additional controls listed in Brooks and Kurtz (2007, 710) but not in Jensen's original article. The results remain the same.

[10] Specifically, we re-sample 1,000 times with replacement from the distribution of first-differences for the respective quantity of interest (DEI, IEI, or TEI) and record the mean of each such bootstrap. This results in a distribution of 1,000 bootstrapped means, from which we construct the 95% confidence interval through normal approximation. This is implemented through R's `boot` package.

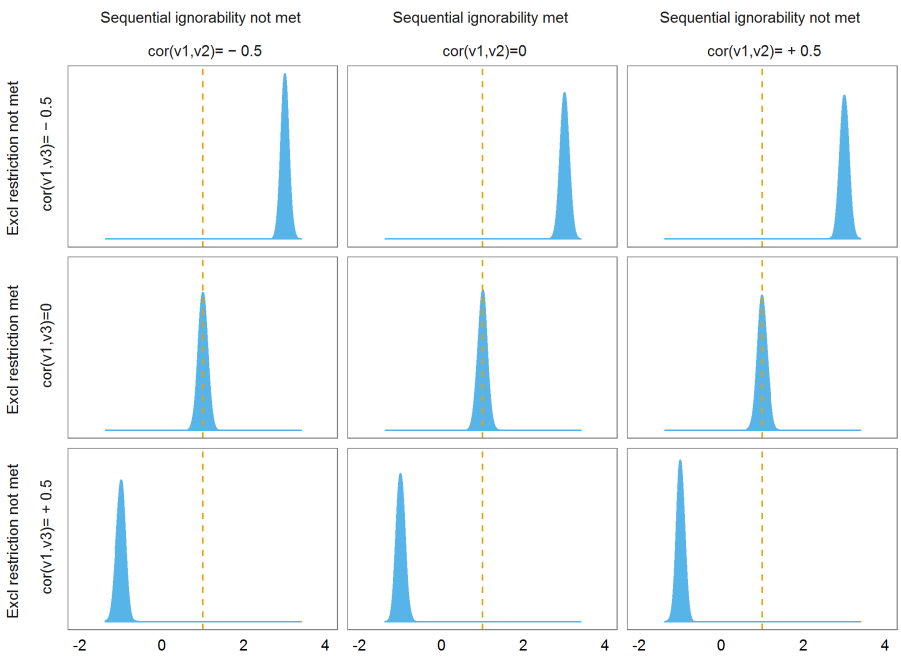
These results, to us, suggest that the effect of democracy on FDI does *not* primarily operate through more liberal policies. Instead, our findings are consistent with an interpretation based on audience costs and increased credibility through democratic institutions. This being said, remember that the DEI is the sum of all effects that are not modeled in the IEI. As such, the DEI may be a combination of various factors. We therefore did not demonstrate that audience costs are the sole mechanism through which democracy operates. But in the spirit of Popper, we showed that we can rule out the competing hypothesis that relies on democracies adopting the right kind of liberal policies.

6 Conclusion

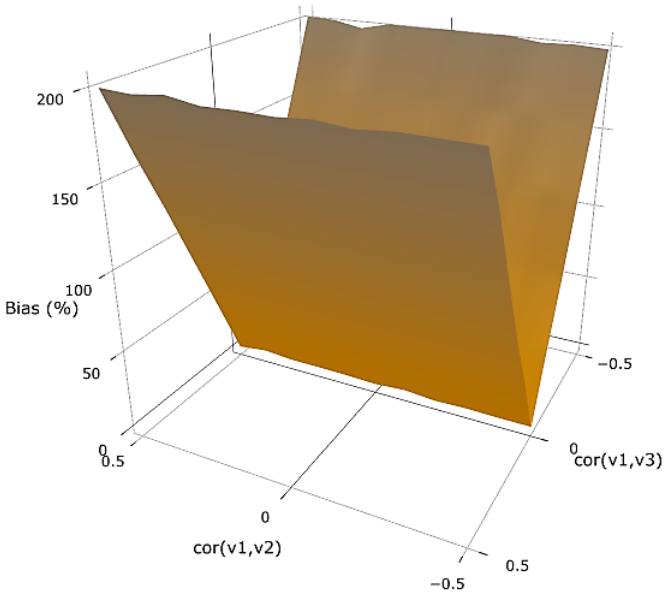
Disentangling the different ways in which causal processes can operate is an important part of social science research. This is particularly true when the processes in question are diffuse. For instance, democracy does many things before it affects the outcomes we are interested in. We need appropriate tools to identify the causal channels that truly matter. Unfortunately, applied researchers are often trapped between a rock (post-treatment bias) and a hard place (omitted variable bias).

In this paper, we offered a simple tool to solve this difficult dilemma. Under relatively benign assumptions, we can recover estimates of the direct, indirect, and total effects of an intervention. This creates valuable opportunities for research in other areas. Oftentimes, theories do not compete in terms of their main predictions. Instead, they differ with respect to the causal channels that they believe are more relevant. But as we know, adjudicating between different causal channels remains a major challenge. We believe that the tools provided in this paper will help researchers venture in this difficult terrain.

At the same time, we make it abundantly clear that the quality of the inferences that we can draw with our method is only as good as the theory underpinning the analysis. We heartily support recent calls for tighter overlaps between theory and empirical work (Samii, 2016). In particular, we believe that thinking hard about causal channels and opening up the "black box of causality" is essential (Imai et al., 2011).

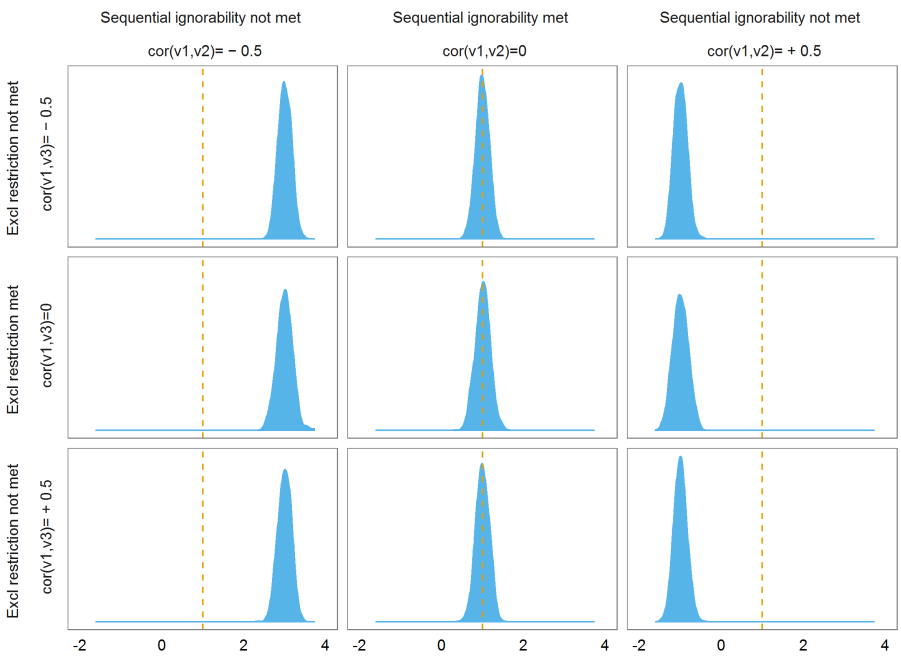


(a) Kernel density of DEI estimates.

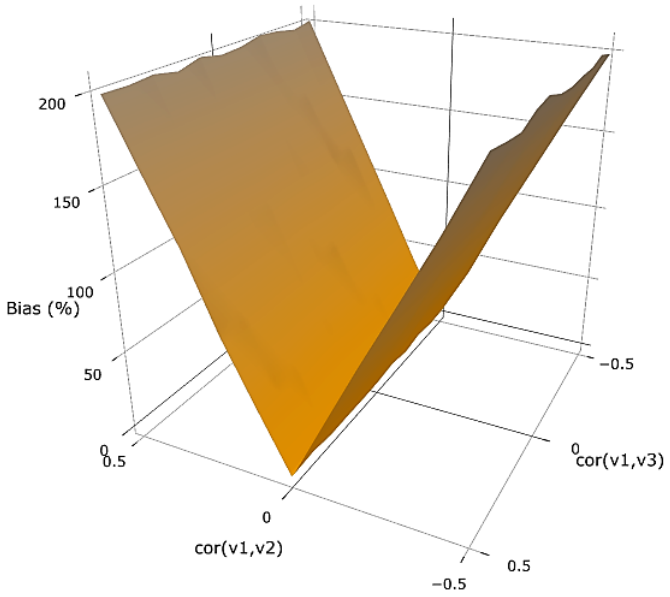


(b) 3D heatmap of bias of DEI estimates (in %).

Figure 2: Results of Monte Carlo simulation for IV approach: The top panel shows the kernel density (bandwidth=0.05) of direct effect estimates for different error correlations. The *true* value of the DEI is 1 (dashed line), and an unbiased estimate can be retrieved when the exclusion restriction assumption holds ($cor(v_1, v_3) = 0$). The bottom panel shows a 3D heatmap of the bias of DEI estimates (in %). Darker areas indicate higher bias; an interactive graph is available online at https://goo.gl/B852Qz.

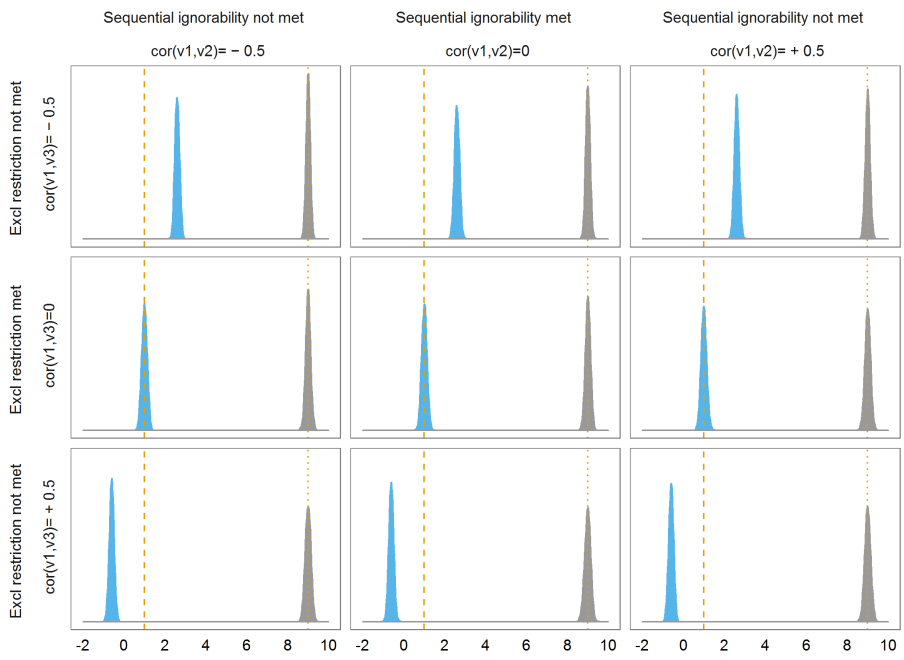


(a) Kernel density of DEI estimates.

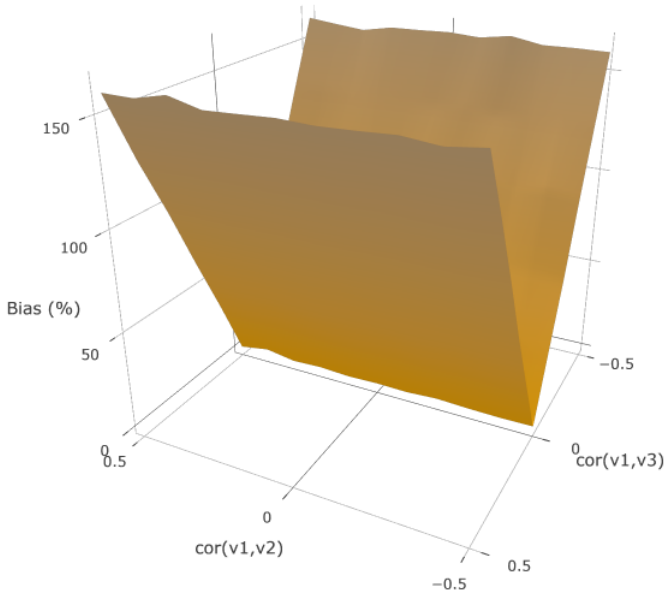


(b) 3D heatmap of bias of DEI estimates (in %).

Figure 3: Results of Monte Carlo simulation for ACDE approach: The top panel shows the kernel density (bandwidth=0.05) of direct effect estimates for different error correlations. The *true* value of the DEI is 1 (dashed line), and an unbiased estimate can be retrieved when the sequential ignorability assumption holds ($cor(v_1, v_2) = 0$). The bottom panel shows a 3D heatmap of the bias of DEI estimates (in %). Darker areas indicate higher bias; an interactive graph is available online at https://goo.gl/1sHWJO.



(a) Kernel density of DEI (blue) and TEI (gray) estimates.



(b) 3D heatmap of bias of DEI estimates (in %).

Figure 4: Results of Monte Carlo simulation for TED approach: The top panel shows the kernel density (bandwidth=0.05) of direct (blue) and total effect (gray) estimates for different error correlations. The *true* value of the DEI is 1 (dashed line) and of the TEI is 9 (dotted line), and an unbiased estimate can be retrieved when a relaxed exclusion restriction assumption holds ($cor(v_1, v_3)_{D=0} = 0$). The bottom panel shows a 3D heatmap of the bias of DEI estimates (in %). Darker areas indicate higher bias; an interactive graph is available online at https://goo.gl/Tsoisf.

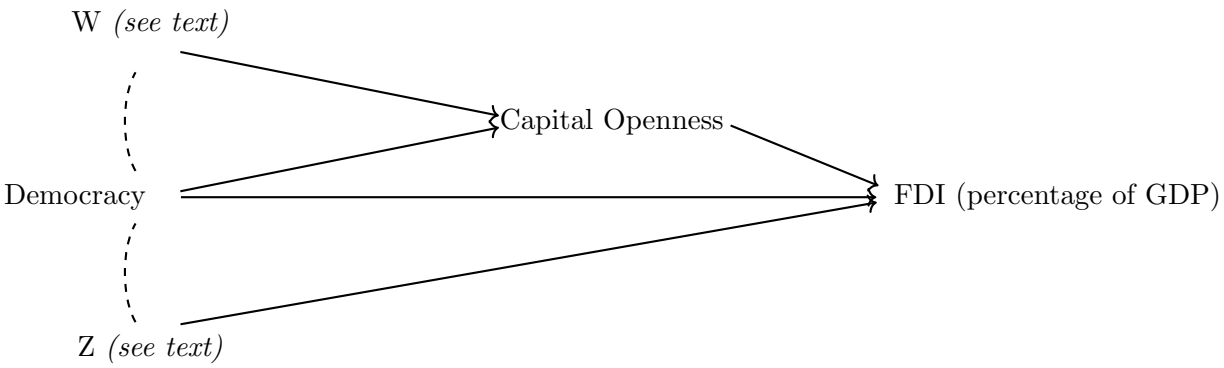


Figure 5: Causal path of democracy on FDI inflows. Adapted from Jensen (2003).

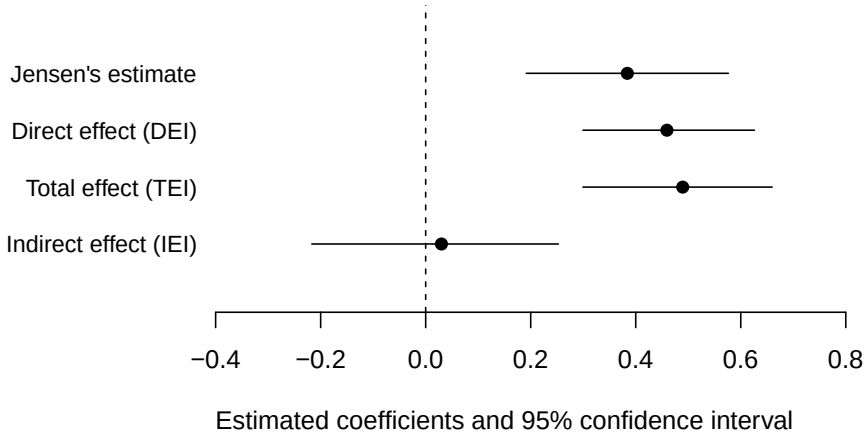


Figure 6: Comparison of Jensen's estimated effect of democracy on FDI inflows and estimated total, direct, and indirect effects from our proposed method. 95% confidence intervals for TEI, DEI, and IEI come from nonparametric bootstrapping.

References

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. 2010. "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program." *Journal of the American Statistical Association* 105 (490): 493–505.

Acharya, Avidit, Matthew Blackwell, and Maya Sen. 2016. "Explaining Causal Findings Without Bias: Detecting and Assessing Direct Effects." *American Political Science Review* 110 (3): 512–529.

Alvarez, Michael R., José Antonio Cheibub, Fernando Limongi, and Adam Przeworski. 1996. "Classifying Political Regimes." *Studies in Comparative International Development* 31 (2): 3–36.

Angrist, Joshua, and Steffan J. Pischke. 2008. *Mostly Harmless Econometics*. Princeton: Princeton University Press.

Brooks, Sarah M., and Marcus J. Kurtz. 2007. "Capital, Trade, and the Political Economies of Reform." *American Journal of Political Science* 51 (4): 703–720.

Chaudoin, Stephen. 2014. "Promises or Policies? An Experimental Analysis of International Agreements and Audience Reactions." *International Organization* 68 (1): 235–256.

Chowdhury, Abdur, and George Mavrotas. 2006. "FDI and Growth: What Causes What?" *World Economy* 29 (1): 9–19.

de Mello, Luiz R. 1997. "Foreign Direct Investment in Developing Countries and Growth: A Selective Survey." *Journal of Development Studies* 34 (1): 1–34.

de Mello, Luiz R. 1999. "Foreign Direct Investment-Led Growth: Evidence From Time Series and Panel Data." *Oxford Economic Papers* 51 (1): 133–151.

Frye, Timothy, and Edward D. Mansfield. 2004. "Timing is Everything: Elections and Trade Liberalization in the Postcommunist World." *Comparative Political Studies* 37 (4): 371–398.

Gerber, Alan S., and Donald P. Green. 2012. *Field Experiments: Design, Analysis, and Interpretation*. New York: W.W. Norton.

Greene, William H. 2008. *Econometric Analysis*. 6th ed. New York: Pearson.

Ho, Daniel E, Kosuke Imai, Gary King, and Elizabeth A. Stuart. 2007. "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." *Political Analysis* 15 (3): 199–236.

Iacus, Stefano M., Gary King, and Giuseppe Porro. 2012. "Causal Inference without Balance Checking: Coarsened Exact Matching." *Politica* 20 (1): 1–24.

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 (4): 765–789.

Jensen, Nathan M. 2003. "Democratic Governance and Multinational Corporations: Political Regimes and Inflows of Foreign Direct Investment." *International Organization* 57 (3): 587–616.

Karcher, Sebastian, and David A Steinberg. 2013. "Assessing the Causes of Capital Account Liberalization: How Measurement Matters." *International Studies Quarterly* 57 (1): 128–137.

King, Gary. 2010. "A Hard Unsolved Problem? Post-treatment Bias in Big Social Science Questions." Presented at the "Hard Problem in Social Science" Symposium, Harvard University.

King, Robert G., and Ross Levine. 1993. "Finance and Growth: Schumpeter Might be Right." *Quarterly Journal of Economics* 108 (3): 717–737.

Kobrin, Stephen J. 1987. "Testing the Bargaining Hypothesis in the Manufacturing Sector in Developing Countries." *International Organization* 41 (4): 609–638.

Milner, Helen V., and Keiko Kubota. 2005. "Why the Move to Free Trade? Democracy and Trade Policy in the Developing Countries." *International Organization* 59 (1): 107–143.

Montgomery, Jacob M., Brendan Nyhan, and Michelle Torres. 2016. "How Conditioning on Post-Treatment Variables Can Ruin Your Experiment and What To Do About It." Working Paper.

Rosenbaum, Paul R. 1984. "The Consequences of Adjustment for a Concomitant Variable that has been Affected by the Treatment." *Journal of the Royal Statistical Society* 147 (5): 656–666.

Samii, Cyrus. 2016. "Causal Empiricism in Quantitative Research." *Journal of Politics* 78 (3): 941–955.