

Estimating and Bounding Mechanism Specific Causal Effects ^{*}

Adam N. Glynn[†]

July 3, 2008

Abstract

Political scientists often cite the importance of mechanism specific causal knowledge, both for its intrinsic scientific value and as a necessity for informed policy. However, outside the framework of additive linear regression models with homogenous causal effects, mechanism specific effects are, in general, not estimated explicitly. Counterfactual causal models allow the formal definition of such concepts as direct, indirect, and mechanism specific effects, and the derivation of conditions for their identification (point or interval). In this paper, I demonstrate the use of counterfactuals to decompose causal effects into mechanism specific effects, showing that estimation and bounding can be accomplished with minor adjustments to standard techniques. I illustrate this methodology with examples from American and Comparative Politics.

^{*}The author thanks Kevin Quinn, Gary King, and Nahomi Ichino. The usual caveat applies.

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

1 Introduction

Political scientists often cite the importance of mechanism specific causal knowledge, both for its intrinsic scientific value and as a necessity for informed policy. This explicit or implicit focus on causal mechanisms pervades important topics across all empirical subfields of the discipline:

We wish to account for a single behavior at a fixed point in time. But it is behavior that stems from a multitude of prior factors. We can visualize the chain of events with which we wish to deal as contained in a funnel of causality. (Campbell et al., 1960, p. 24)

I do not model presidential election rules as having a direct impact on the legislative party system. Instead, there is a two-step process: (1) Presidential election rules combine interactively with social diversity to produce an effective number of presidential candidates; (2) the effective number of presidential candidates affects the effective number of legislative competitors... (Cox, 1997, p. 204)

The political and military technology of insurgency will be favored, and thus civil war made more likely, when potential rebels face or have available the following ... A state whose revenues derive primarily from oil exports. Oil producers tend to have weaker state apparatuses than one would expect given their level of income because the rulers have less need for a socially intrusive and elaborate bureaucratic system... (Fearon and Laitin, 2003, p. 81)

I adopt two strategies for testing the persuasiveness of the causal logics that underpin democratic peace theory. First, I take each logic at face value and ask whether the hypothesized causal mechanisms operate as stipulated by the theory's proponents... (Rosato, 2003, p. 585)

However, outside the framework of additive linear regression models with homogenous causal effects, mechanism specific effects are, in general, not estimated explicitly. In fact, many authors have noticed that currently utilized quantitative political methods cannot be used to investigate causal mechanisms under traditional theoretical assumptions (Collier and Brady, 2004; Hall, 2003). The following quote nicely summarizes the state of the discipline.

What causal mechanisms produce these outcomes? ... Case studies tend to provide a wealth of data on causal links..., but are difficult to generalize. There are plausible theories behind each of the patterns, though efforts to test them are still in their infancy. (Ross, 2004, p. 338)

In this manuscript, I address this methodological deficiency by adapting and expanding the work of Robins and Greenland (1992); Pearl (2001); Robins (2003); Petersen et al. (2006) to show that quantitative methods *can* be used to estimate (or at least bound) mechanism specific effects. Furthermore, this can be accomplished outside of the additive linear model framework—allowing interactions, non-linear models, and causal effect heterogeneity.

This paper is organized as follows. In Section 2, I present a counterfactual definition of causal mechanisms, define mechanism specific effects, and discuss learning about individual level mechanism specific effects from observational data. In Section 3, I present conditions for point and interval identification of average mechanism specific effects, showing in particular that traditional ignorability assumptions (and hence traditional experimental designs) are not sufficient for the point identification of average mechanism specific effects. In Section 4, I present illustrative applications of estimation and bounding for mechanism specific effects. Section 5 concludes.

2 A Counterfactual Definition of Mechanisms

Following Rubin (1974), Robins (1986), and Pearl (2000) I define a recursive counterfactual causal model over a set of causally ordered (indexed by k) measured variables $\{V_{1i}, V_{2i}, \dots, V_{ki}, \dots, V_{Ki}\}$ for units $i = 1, \dots, n$. For this model, I assume that there is no interference between units and that the following counterfactual variables are well defined:

Definition 1 (Potential Variables) *Potential variable values are written as the following (many of these variables will be counterfactual):*

- a) $V_{ki}(V_{ji} = v)$ is the value of V_{ki} that unit i would have had if V_{ji} had been v . If $j > k$, then this is equivalent to the observed V_{ki} due to causal order.
- b) $V_{ki}(V_{ji} = v_j, V_{j+1i} = v_{j+1}, \dots, V_{j+pi} = v_{j+p})$ is the value of V_{ki} that unit i would have had if V_{ji} had been v_j , and V_{j+1i} had been v_{j+1} , and \dots V_{j+pi} had been v_{j+p} . Again, if $j > k$, then this is equivalent to the observed V_{ki} due to causal order.

It is often efficient to represent the causal order over these variables with a Directed Acyclic Graph (DAG). In Figure 1, I present a DAG for a causal model with four measured variables. In this case the variable U represents a vector of all the unobserved background factors that affect the four measured variables, and the arrows (more properly missing arrows) in the graph represent the causal order assumption.

Within this framework, we can loosely define the causal effect of V_j on V_k to be represented by all directed sequences of arrows that lead from V_j to V_k , and we can further consider decomposing this effect into mechanism specific effects represented by each specific sequence. For example, in Figure

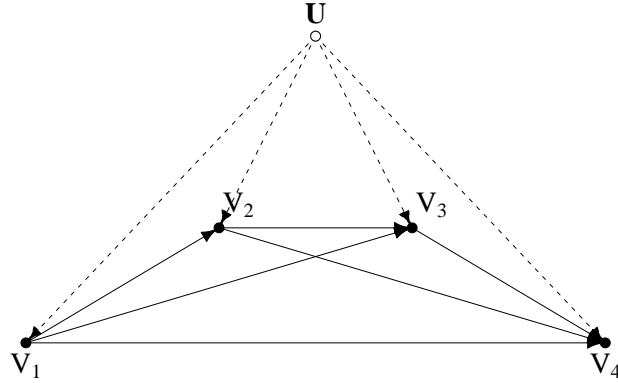


Figure 1: *Directed Acyclic Graph (DAG) consistent with a recursive counterfactual causal model with four measured variables.*

1, we might think of the causal effect of V_1 on V_4 as being composed of the following mechanism specific effects:

$$V_1 \rightarrow V_2 \rightarrow V_3 \rightarrow V_4$$

$$V_1 \rightarrow V_2 \rightarrow V_4$$

$$V_1 \rightarrow V_3 \rightarrow V_4$$

$$V_1 \rightarrow V_4.$$

For example, the first mechanism specified above might be interpreted as the effect of V_1 on V_4 that “goes through” V_2 and V_3 . However, there are subtleties in this definition and the identification of mechanism specific effects that will not be readily apparent from the graph or the intuitive interpretation. In the remainder of this paper, I focus solely on models with three measured variables in order to explicate these subtleties.

2.1 Counterfactual Causal Mechanisms with Three Measured Variables

Consider an example with three measured variables $\{X, Z, Y\}$ for individuals indexed by $i = 1, \dots, n$, where X_i represents a pre-existing explanatory variable, Z_i represents an explanatory vari-

able that may or may not be affected by the variable X_i (Z_i is sometimes called a concomitant variable), and Y_i represents an outcome variable which may or may not be affected by X_i and Z_i . This model is represented by Figure 2.

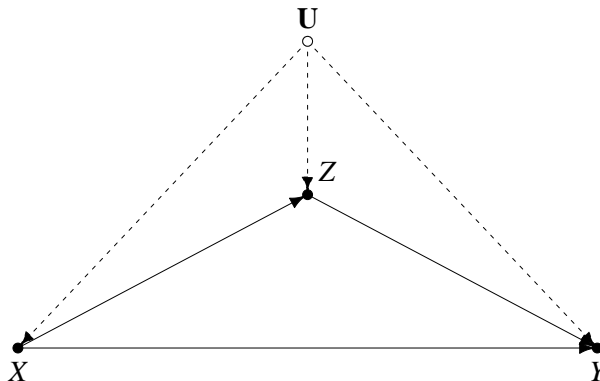


Figure 2: *Directed Acyclic Graph (DAG) consistent with a recursive counterfactual causal model with three measured variables.*

With the causal order specified in Figure 2 we can define a number of potential variables for each individual:

Definition 2 (Potential Variables in the Three Variable Model) *The potential variables in this model can be represented with the following notation:*

- a) $Y_i(X_i = x)$ is the potential outcome that individual i would have if they had the value x for the variable X_i .
- b) $Y_i(Z_i = z)$ is the potential outcome that individual i would have if they had the value z for the variable Z_i .
- c) $Z_i(X_i = x)$ is the potential concomitant value that individual i would have if they had the value x for the variable X_i .
- d) $Y_i(X_i = x, Z_i = z)$ is the potential outcome that individual i would have if they had the value the value x for the variable X_i and z for the variable Z_i .

Depending on the observed values for X and Z , some of these potential variables will be observed, while some will be counterfactual. As an example, we consider the case where all three variables X , Y , and Z are binary. Table 1 presents the possible values for the potential variables in

Table 1: Possible observed values for the potential variables in Definition 2 a), b), and c) when X_i , Z_i , and Y_i are binary. For many of the potential variables, we only observe a set of possible values.

X_i	Z_i	Y_i	$Y_i(X_i = 0)$	$Y_i(X_i = 1)$	$Z_i(X_i = 0)$	$Z_i(X_i = 1)$	$Y_i(Z_i = 0)$	$Y_i(Z_i = 1)$
0	0	0	0	{0, 1}	0	{0, 1}	0	{0, 1}
0	0	1	1	{0, 1}	0	{0, 1}	1	{0, 1}
0	1	0	0	{0, 1}	1	{0, 1}	{0, 1}	0
0	1	1	1	{0, 1}	1	{0, 1}	{0, 1}	1
1	0	0	{0, 1}	0	{0, 1}	0	0	{0, 1}
1	0	1	{0, 1}	1	{0, 1}	0	1	{0, 1}
1	1	0	{0, 1}	0	{0, 1}	1	{0, 1}	0
1	1	1	{0, 1}	1	{0, 1}	1	{0, 1}	1

Table 2: Possible observed values for the potential variables in Definition 2 d) when X_i , Z_i , and Y_i are binary. For many of the potential variables, we only observe a set of possible values.

X_i	Z_i	Y_i	$Y_i(X_i = 0, Z_i = 0)$	$Y_i(X_i = 1, Z_i = 0)$	$Y_i(X_i = 0, Z_i = 1)$	$Y_i(X_i = 1, Z_i = 1)$
0	0	0	0	{0, 1}	{0, 1}	{0, 1}
0	0	1	1	{0, 1}	{0, 1}	{0, 1}
0	1	0	{0, 1}	{0, 1}	0	{0, 1}
0	1	1	{0, 1}	{0, 1}	1	{0, 1}
1	0	0	{0, 1}	0	{0, 1}	{0, 1}
1	0	1	{0, 1}	1	{0, 1}	{0, 1}
1	1	0	{0, 1}	{0, 1}	{0, 1}	0
1	1	1	{0, 1}	{0, 1}	{0, 1}	1

Definition 2 a), b) and c) in this scenario. For some values of X , Y , and Z the potential variables are observed while in other cases, we only observe a set of possible values. Table 2 presents the possible values for the potential variables in Definition 2 d). Again, for some values of X , Y , and Z these joint potential variables are observed while in other cases, we only observe a set of possible values. However, note that in this example, we only observe one of the four joint potential values for each individual.

Using the potential variables from Definition 2 a), b), and c), we can define a number of individual causal effects as in Holland (1986).

Definition 3 (Individual Total Effects) *The following contrasts are referred to as total effects, because they represent the overall effect of changing a single causal variable.*

- a) $Y_i(X_i = x') - Y_i(X_i = x)$ is the difference in outcome that individual i would have shown when comparing the outcome they would have had if X_i had been x to the outcome they would have had if X_i had been x' .

Table 3: Possible values for the individual total effects in Definition 3 when X_i , Z_i , and Y_i are binary.

X_i	Z_i	Y_i	$Y_i(X_i = 1) - Y_i(X_i = 0)$	$Y_i(Z_i = 1) - Y_i(Z_i = 0)$	$Z_i(X_i = 1) - Z_i(X_i = 0)$
0	0	0	{0, 1}	{0, 1}	{0, 1}
0	0	1	{-1, 0}	{-1, 0}	{0, 1}
0	1	0	{0, 1}	{-1, 0}	{-1, 0}
0	1	1	{-1, 0}	{0, 1}	{-1, 0}
1	0	0	{-1, 0}	{0, 1}	{-1, 0}
1	0	1	{0, 1}	{-1, 0}	{-1, 0}
1	1	0	{-1, 0}	{-1, 0}	{0, 1}
1	1	1	{0, 1}	{0, 1}	{0, 1}

- b) $Y_i(Z_i = z') - Y_i(Z_i = z)$ is the difference in outcome that individual i would have shown when comparing the outcome they would have had if Z_i had been z to the outcome they would have had if Z_i had been z' .
- c) $Z_i(X_i = x') - Z_i(X_i = x)$ is the difference in the concomitant that individual i would have shown when comparing the concomitant they would have had if X_i had been x to the concomitant they would have had if X_i had been x' .

Notice that while these individual total effects can never be observed (due to the fact that we can observe at most one of the values within each effect), when we have bounds on the support of Z and Y , we can learn *something* about these effects from observational data. In Table 3, I present the possible values for the effects in Definition 3 when X , Y , and Z are binary. In this case, all of these effects can only take on values in the set $\{-1, 0, 1\}$. However, we see in the table that for each of these effects we observe at least one of the potential variable values, and therefore the set of possibilities is reduced by one element.

In addition to the traditional causal effects in Definition 3, we can use the potential variables from Definition 2 d), to define additional individual causal effects as in Pearl (2001).

Definition 4 (Individual Controlled Direct Effects) *The following contrasts are referred to as controlled direct effects, because they represent the effect of changing a single causal variable while another causal variable is held at a constant value.*

- a) $Y_i(X_i = x', Z_i = z) - Y_i(X_i = x, Z_i = z)$ is the difference in outcome that individual i would have shown when comparing the outcome they would have had if X_i had been x and Z_i had been z to the outcome they would have had if X_i had been x' and Z_i had been z .
- b) $Y_i(X_i = x, Z_i = z') - Y_i(X_i = x, Z_i = z)$ is the difference in outcome that individual i would have shown when comparing the outcome they would have had if X_i had been x and Z_i had been z to the outcome they would have had if X_i had been x and Z_i had been z' .

Table 4: Possible values for the individual controlled direct effects in Definition 4 when X_i , Z_i , and Y_i are binary.

X_i	Z_i	Y_i	$Y_i(X_i = 1, Z_i = 0) - Y_i(X_i = 0, Z_i = 0)$	$Y_i(X_i = 1, Z_i = 1) - Y_i(X_i = 0, Z_i = 1)$
0	0	0	{0, 1}	{-1, 0, 1}
0	0	1	{-1, 0}	{-1, 0, 1}
0	1	0	{-1, 0, 1}	{0, 1}
0	1	1	{-1, 0, 1}	{-1, 0}
1	0	0	{-1, 0}	{-1, 0, 1}
1	0	1	{0, 1}	{-1, 0, 1}
1	1	0	{-1, 0, 1}	{-1, 0}
1	1	1	{-1, 0, 1}	{0, 1}
X_i	Z_i	Y_i	$Y_i(X_i = 0, Z_i = 1) - Y_i(X_i = 0, Z_i = 0)$	$Y_i(X_i = 1, Z_i = 1) - Y_i(X_i = 1, Z_i = 0)$
0	0	0	{0, 1}	{-1, 0, 1}
0	0	1	{-1, 0}	{-1, 0, 1}
0	1	0	{-1, 0}	{-1, 0, 1}
0	1	1	{0, 1}	{-1, 0, 1}
1	0	0	{-1, 0, 1}	{0, 1}
1	0	1	{-1, 0, 1}	{-1, 0}
1	1	0	{-1, 0, 1}	{-1, 0}
1	1	1	{-1, 0, 1}	{0, 1}

Again, these individual controlled direct effects cannot be observed (due to the fact that we can observe at most one of the values within each effect), however, when we have bounds on the support of Z and Y , we can again learn *something* about these effects from observational data. In Table 4, I present the possible values for the effects in Definition 4 when X , Y , and Z are binary. Again these effects can only logically take on values in the set $\{-1, 0, 1\}$, however, we see in the table that only for some combinations of the observed variables will the set of possibilities is reduced.

While individual total or controlled direct effects and/or sample or population summaries of these effects may be of interest, in order to define indirect effects, or to decompose total effects into mechanism specific effects, we must define more complex counterfactual quantities. To do this note that we can write the potential outcome $Y_i(X_i = x)$ in a redundant manner as combinations of the potential variables in Definition 2 c) and d):

$$Y_i(X_i = x) = Y_i(X_i = x, Z_i = Z_i(X_i = x)) \quad (1)$$

where $Y_i(X_i = x, Z_i = Z_i(X_i = x))$ indicates the outcome that individual i would have had if their

X_i value had been x and if their Z_i value had been what it would have been if their X_i value had been x . This is clearly just a long winded way of saying $Y_i(X_i = x)$ indicates the outcome that individual i would have had if their X_i value had been x , however, this type of expression can be used to define more complex counterfactuals that allow a formalization of causal mechanisms:

Definition 5 (Complex Counterfactual Variables) *Complex counterfactual variables are counterfactual variables that require the simultaneous consideration of different values for a single treatment variable. For example, $Y_i(X_i = x, Z_i = Z_i(X_i = x'))$ indicates the outcome that individual i would have had if their X_i value had been x and if their Z_i value had been what it would have been if their X_i value had been x' . When Z is discrete, this quantity can be written as a combination of potential variables as introduced in Definition 2:*

$$Y_i(X_i = x, Z_i = Z_i(X_i = x')) = \sum_z Y_i(X_i = x, Z_i = z) 1_{\{Z_i(X_i=x')=z\}}$$

where $1_{\{\cdot\}}$ is an indicator function.

The decomposition in Definition 5 is easier to contemplate when the variables X , Y , and Z are binary. In this case,

$$\begin{aligned} Y_i(X_i = 0, Z_i = Z_i(X_i = 1)) &= Y_i(X_i = 0, Z_i = 0) \cdot (1 - Z_i(X_i = 1)) \\ &\quad + Y_i(X_i = 0, Z_i = 1) \cdot Z_i(X_i = 1) \end{aligned} \quad (2)$$

$$\begin{aligned} Y_i(X_i = 1, Z_i = Z_i(X_i = 0)) &= Y_i(X_i = 1, Z_i = 0) \cdot (1 - Z_i(X_i = 0)) \\ &\quad + Y_i(X_i = 1, Z_i = 1) \cdot Z_i(X_i = 0), \end{aligned} \quad (3)$$

where these quantities were all defined in Tables 1 and 2. Furthermore, using the redundant potential outcomes notation in (1) and the complex counterfactuals in Definition 5, the traditional individual total causal effects of Definition 3 can be decomposed into mechanism specific effects by adding and subtracting complex counterfactuals:

Definition 6 (Individual Mechanism Specific Effects) *The Individual Total Effect on Y of changing X from x to x' can be decomposed into mechanism specific effects in the following two ways:*

a)

$$\begin{aligned} Y_i(X_i = x') - Y_i(X_i = x) &= Y_i(X_i = x', Z_i = Z_i(X_i = x')) - Y_i(X_i = x, Z_i = Z_i(X_i = x)) \\ &= Y_i(X_i = x', Z_i = Z_i(X_i = x')) - Y_i(X_i = x, Z_i = Z_i(X_i = x')) \\ &\quad + Y_i(X_i = x, Z_i = Z_i(X_i = x')) - Y_i(X_i = x, Z_i = Z_i(X_i = x)) \end{aligned}$$

where the last line represents the portion of the total effect that goes indirectly through the variable Z , the second to last line represents the portion of the total effect that goes directly through all unspecified mechanisms.

b)

$$\begin{aligned} Y_i(X_i = x') - Y_i(X_i = x) &= Y_i(X_i = x', Z_i = Z_i(X_i = x')) - Y_i(X_i = x, Z_i = Z_i(X_i = x)) \\ &= Y_i(X_i = x', Z_i = Z_i(X_i = x')) - Y_i(X_i = x', Z_i = Z_i(X_i = x)) \\ &\quad + Y_i(X_i = x', Z_i = Z_i(X_i = x)) - Y_i(X_i = x, Z_i = Z_i(X_i = x)) \end{aligned}$$

where the last line represents the portion of the total effect that goes directly through all unspecified mechanisms, and the second to last line represents the portion of the total effect that goes indirectly through the variable Z .

When X , Y , and Z are binary, the individual total effect $Y_i(X_i = 1) - Y_i(X_i = 0)$ can be decomposed in the following two ways:

$$Y_i(X_i = 1) - Y_i(X_i = 0) = Y_i(X_i = 1, Z_i(X_i = 1)) - Y_i(X_i = 1, Z_i(X_i = 0)) \quad (4)$$

$$+ \{Y_i(X_i = 1, Z_i(X_i = 0)) - Y_i(X_i = 0, Z_i(X_i = 0))\} \quad (5)$$

$$Y_i(X_i = 1) - Y_i(X_i = 0) = Y_i(X_i = 1, Z_i(X_i = 1)) - Y_i(X_i = 0, Z_i(X_i = 1)) \quad (6)$$

$$+ \{Y_i(X_i = 0, Z_i(X_i = 1)) - Y_i(X_i = 0, Z_i(X_i = 0))\}. \quad (7)$$

Here, (4) and (7)¹ represent effects that are specific to the Z mechanism, while (5)² and (6) represent effects that are specific to the remaining mechanisms.

Again, these individual mechanism specific effects cannot be observed (due to the fact that we can observe at most one of the values within each effect), however, when we have bounds on the support of Z and Y , we may be able to learn *something* about these effects from observational data. In Table 5, I present the possible values for the effects in Definition 6 when X , Y , and Z are binary. Again these effects can only logically take on values in the set $\{-1, 0, 1\}$, however, we see in the table that only for some combinations of the observed variables will the set of possibilities be reduced. Furthermore, note that the observed value of Z is irrelevant for these sets. Therefore, by itself, observing Z tell us nothing about the individual level mechanism specific effects.

¹This is sometimes called a pure (Robins and Greenland, 1992) or natural (Pearl, 2001) indirect effect.

²This is sometimes called pure (Robins and Greenland, 1992) or natural (Pearl, 2001) direct effect and will be equivalent to principal stratification direct effects (Frangakis and Rubin, 2002) when $Z_i(X_i = 1) = Z_i(X_i = 0)$.

Table 5: Possible values for the individual mechanism specific effects in Definition 5 when X_i , Z_i , and Y_i are binary.

X_i	Z_i	Y_i	$Y_i(X_i = 1, Z_i(X_i = 1)) - Y_i(X_i = 1, Z_i(X_i = 0))$	$Y_i(X_i = 1, Z_i(X_i = 0)) - Y_i(X_i = 0, Z_i(X_i = 0))$
0	0	0	$\{-1, 0, 1\}$	$\{0, 1\}$
0	0	1	$\{-1, 0, 1\}$	$\{-1, 0\}$
0	1	0	$\{-1, 0, 1\}$	$\{0, 1\}$
0	1	1	$\{-1, 0, 1\}$	$\{-1, 0\}$
1	0	0	$\{-1, 0\}$	$\{-1, 0, 1\}$
1	0	1	$\{0, 1\}$	$\{-1, 0, 1\}$
1	1	0	$\{-1, 0\}$	$\{-1, 0, 1\}$
1	1	1	$\{0, 1\}$	$\{-1, 0, 1\}$
X_i	Z_i	Y_i	$Y_i(X_i = 1, Z_i(X_i = 1)) - Y_i(X_i = 0, Z_i(X_i = 1))$	$Y_i(X_i = 0, Z_i(X_i = 1)) - Y_i(X_i = 0, Z_i(X_i = 0))$
0	0	0	$\{-1, 0, 1\}$	$\{0, 1\}$
0	0	1	$\{-1, 0, 1\}$	$\{-1, 0\}$
0	1	0	$\{-1, 0, 1\}$	$\{0, 1\}$
0	1	1	$\{-1, 0, 1\}$	$\{-1, 0\}$
1	0	0	$\{-1, 0\}$	$\{-1, 0, 1\}$
1	0	1	$\{0, 1\}$	$\{-1, 0, 1\}$
1	1	0	$\{-1, 0\}$	$\{-1, 0, 1\}$
1	1	1	$\{0, 1\}$	$\{-1, 0, 1\}$

2.2 Learning About Individual Mechanism Specific Effects

As shown in Table 5, the observation of intermediate variables by itself, does not improve our causal knowledge about individual level total or mechanism specific effects. However, this analysis neglects the fact that intermediate variables can be chosen so as to justify monotonicity assumptions and exclusion restrictions. In this section, I demonstrate the utility of such assumptions for learning about mechanism specific effects. An analysis with binary X , Y , and Z will suffice to demonstrate the most important points.

Suppose that Z is chosen so that the total effect of X on Z will never be negative:

$$Z_i(X_i = 1) - Z_i(X_i = 0) \geq 0 \text{ for all } i. \quad (8)$$

If (8) holds, and we observe $X_i = 1$ and $Z_i = 0$, then we know that $Z_i(X_i = 0) = Z_i(X_i = 1) = 0$.

If (8) holds, and we observe $X_i = 0$ and $Z_i = 1$, then we know that $Z_i(X_i = 1) = Z_i(X_i = 0) = 1$. Furthermore, notice that due to these relationships and the decomposition of total effects into mechanism specific effects, when the assumption (8) holds, the set of possible values for the

Table 6: Possible values for the individual mechanism specific effects in Definition 6 when X_i , Z_i , and Y_i are binary. Values in red are not possible when the monotonicity assumption in (8) holds.

X_i	Z_i	Y_i	$Y_i(X_i = 1, Z_i(X_i = 1)) - Y_i(X_i = 1, Z_i(X_i = 0))$	$Y_i(X_i = 1, Z_i(X_i = 0)) - Y_i(X_i = 0, Z_i(X_i = 0))$
0	0	0	$\{-1, 0, 1\}$	$\{0, 1\}$
0	0	1	$\{-1, 0, 1\}$	$\{-1, 0\}$
0	1	0	$\{-1, 0, 1\}$	$\{0, 1\}$
0	1	1	$\{-1, 0, 1\}$	$\{-1, 0\}$
1	0	0	$\{-1, 0\}$	$\{-1, 0, 1\}$
1	0	1	$\{0, 1\}$	$\{-1, 0, 1\}$
1	1	0	$\{-1, 0\}$	$\{-1, 0, 1\}$
1	1	1	$\{0, 1\}$	$\{-1, 0, 1\}$
X_i	Z_i	Y_i	$Y_i(X_i = 1, Z_i(X_i = 1)) - Y_i(X_i = 0, Z_i(X_i = 1))$	$Y_i(X_i = 0, Z_i(X_i = 1)) - Y_i(X_i = 0, Z_i(X_i = 0))$
0	0	0	$\{-1, 0, 1\}$	$\{0, 1\}$
0	0	1	$\{-1, 0, 1\}$	$\{-1, 0\}$
0	1	0	$\{-1, 0, 1\}$	$\{0, 1\}$
0	1	1	$\{-1, 0, 1\}$	$\{-1, 0\}$
1	0	0	$\{-1, 0\}$	$\{-1, 0, 1\}$
1	0	1	$\{0, 1\}$	$\{-1, 0, 1\}$
1	1	0	$\{-1, 0\}$	$\{-1, 0, 1\}$
1	1	1	$\{0, 1\}$	$\{-1, 0, 1\}$

mechanism specific effects will be reduced for some combinations of observed X , Y , and Z . In Table 6, I have re-printed Table 5 with the values that are impossible under assumption (8) printed in red.

Alternatively, suppose that Z is chosen so that the controlled direct effect of Z on Y is never negative:

$$Y_i(X_i = 0, Z_i = 1) - Y_i(X_i = 0, Z_i = 0) \geq 0 \text{ for all } i. \quad (9)$$

$$Y_i(X_i = 1, Z_i = 1) - Y_i(X_i = 1, Z_i = 0) \geq 0 \text{ for all } i \quad (10)$$

When (9) holds, then due to the decomposition in (2), we know that $Y_i(X_i = 0, Z_i = Z_i(X_i = 1)) = 1$ when we observe $X_i = 0$, $Z_i = 0$, and $Y_i = 1$, and we know that $Y_i(X_i = 0, Z_i = Z_i(X_i = 1)) = 0$ when we observe $X_i = 0$, $Z_i = 1$, and $Y_i = 0$. Similarly, when (10) holds, then due to the decomposition in (3), we know that $Y_i(X_i = 1, Z_i = Z_i(X_i = 0)) = 1$ when we observe $X_i = 1$, $Z_i = 0$, and $Y_i = 1$, and we know that $Y_i(X_i = 1, Z_i = Z_i(X_i = 0)) = 0$ when we observe $X_i = 1$, $Z_i = 1$, and $Y_i = 0$. Furthermore, notice that due to these relationships and the decomposition of total effects into mechanism specific effects, when the assumptions (9) and (10) hold, the set of

Table 7: Possible values for the individual mechanism specific effects in Definition 6 when X_i , Z_i , and Y_i are binary. Values in red are not possible when the monotonicity assumptions in (9) and (10) hold.

X_i	Z_i	Y_i	$Y_i(X_i = 1, Z_i(X_i = 1)) - Y_i(X_i = 1, Z_i(X_i = 0))$	$Y_i(X_i = 1, Z_i(X_i = 0)) - Y_i(X_i = 0, Z_i(X_i = 0))$
0	0	0	$\{-1, 0, 1\}$	$\{0, 1\}$
0	0	1	$\{-1, 0, 1\}$	$\{-1, 0\}$
0	1	0	$\{-1, 0, 1\}$	$\{0, 1\}$
0	1	1	$\{-1, 0, 1\}$	$\{-1, 0\}$
1	0	0	$\{-1, 0\}$	$\{-1, 0, 1\}$
1	0	1	$\{0, 1\}$	$\{-1, 0, 1\}$
1	1	0	$\{-1, 0\}$	$\{-1, 0, 1\}$
1	1	1	$\{0, 1\}$	$\{-1, 0, 1\}$
X_i	Z_i	Y_i	$Y_i(X_i = 1, Z_i(X_i = 1)) - Y_i(X_i = 0, Z_i(X_i = 1))$	$Y_i(X_i = 0, Z_i(X_i = 1)) - Y_i(X_i = 0, Z_i(X_i = 0))$
0	0	0	$\{-1, 0, 1\}$	$\{0, 1\}$
0	0	1	$\{-1, 0, 1\}$	$\{-1, 0\}$
0	1	0	$\{-1, 0, 1\}$	$\{0, 1\}$
0	1	1	$\{-1, 0, 1\}$	$\{-1, 0\}$
1	0	0	$\{-1, 0\}$	$\{-1, 0, 1\}$
1	0	1	$\{0, 1\}$	$\{-1, 0, 1\}$
1	1	0	$\{-1, 0\}$	$\{-1, 0, 1\}$
1	1	1	$\{0, 1\}$	$\{-1, 0, 1\}$

possible values for the mechanism specific effects will be reduced for some combinations of observed X , Y , and Z . In Table 7, I have re-printed Table 5 with the values that are impossible under assumptions (9) and (10) printed in red.

Finally, suppose that Z is chosen so that the effect of X on Y goes entirely through Z :

$$Y_i(X_i = 0, Z_i = 0) = Y_i(X_i = 1, Z_i = 0) \text{ for all } i. \quad (11)$$

$$Y_i(X_i = 0, Z_i = 1) = Y_i(X_i = 1, Z_i = 1) \text{ for all } i \quad (12)$$

when the assumptions (11) and (12) hold, the set of possible values for the mechanism specific effects will be reduced because the “direct effects” will all be zero. Notice as well that when monotonicity assumptions and exclusion restrictions are combined, individual level total effects will be identified for some individuals.

Even with monotonicity assumptions and exclusion restrictions, the individual level mechanism specific effects are only identified for some individuals, and only in the case when this effect is zero. This difficulty is not special to the case of mechanism specific effects (for individual total effects

Holland (1986) calls this the fundamental problem of causal inference). Due to this fundamental lack of information, researchers often resign themselves to estimating summaries over these individual effects. Typically, sample average effects or population average effects are chosen as the target of inference, and ignorability assumptions are used to justify the point identification of these effects. However, the assumptions required for the point identification of mechanism specific effects are typically stronger than those that are needed for the point identification of average total effects. In the next section I discuss the identification (point or interval) of average mechanism specific effects, showing that ignorability assumptions are not sufficient for the point identification of these effects and presenting a variety of assumptions that can be used to provide interval or point identification.

3 The Identification (Point or Interval) of Average Mechanism Specific Effects

3.1 The Insufficiency of Ignorability for Point Identification of Average Mechanism Specific Effects

A number of authors have shown that the averages of individual total effects $E[Y(X = x') - Y(X = x)]$ can be point identified by regression, matching, or weighting techniques when treatment assignment is weakly mean “ignorable”:

$$E[Y(X = x)|\mathbf{W}] = E[Y|X = x, \mathbf{W}] \text{ for all } x \quad (13)$$

(all conditional expectations in this paper are assumed to be well defined). This assumption will hold when X is “as if randomly assigned” within the strata defined by \mathbf{W} . It seems natural to assume that averages over individual mechanism specific effects might be point identified with similar ignorability assumptions, but this is not the case.

Average mechanism specific effects can be defined as averages over a sample or population of the individual mechanism specific effects of Definition 6:

Definition 7 (Average Mechanism Specific Effects) *The average total effects on Y of changing X from x to x' can be decomposed into average mechanism specific effects in the following two ways (I use the expectation operator ($E[\cdot]$) to simultaneously represent sample or population averages):*

a)

$$\begin{aligned} E[Y(X = x') - Y(X = x)] &= E[Y(X = x', Z = Z(X = x')) - Y(X = x, Z = Z(X = x))] \\ &= E[Y(X = x', Z = Z(X = x')) - Y(X = x, Z = Z(X = x'))] \\ &\quad + E[Y(X = x, Z = Z(X = x')) - Y(X = x, Z = Z(X = x))] \end{aligned}$$

b)

$$\begin{aligned} E[Y(X = x') - Y(X = x)] &= E[Y(X = x', Z = Z(X = x')) - Y(X = x, Z = Z(X = x))] \\ &= E[Y(X = x', Z = Z(X = x')) - Y(X = x', Z = Z(X = x))] \\ &\quad + E[Y(X = x, Z = Z(X = x)) - Y(X = x, Z = Z(X = x))] \end{aligned}$$

In order to determine the conditions for identification, it will be helpful to re-write these effects in terms of the decomposition in Definition 5 (See Appendix A for details).³

$$\begin{aligned} E[Y(X = x, Z = Z(X = x')) - Y(X = x, Z = Z(X = x))] \\ &= E[E[Y(X = x, Z = Z(X = x')) - Y(X = x, Z = Z(X = x)) | \mathbf{W}]] \\ &= E \left[\sum_z \{ E[Y(X = x, Z = z) | \mathbf{W}] E[1_{\{Z(X=x')=z\}} - 1_{\{Z(X=x)=z\}} | \mathbf{W}] \right. \\ &\quad \left. + Cov[Y(X = x, Z = z), 1_{\{Z(X=x')=z\}} - 1_{\{Z(X=x)=z\}} | \mathbf{W}] \right] \end{aligned} \quad (14)$$

Therefore, in order to identify the effect in (14), we need the following to hold for all z , and \mathbf{W} :

$$E[Y(X = x, Z = z) | \mathbf{W}] = E[Y | X = x, Z = z, \mathbf{W}] \quad (15)$$

$$E[1_{\{Z(X=x)=z\}} | \mathbf{W}] = E[1_{\{Z=z\}} | X = x, \mathbf{W}] \quad (16)$$

and we need the following to hold for all \mathbf{W} :

$$\sum_z Cov[Y(X = x, Z = z), 1_{\{Z(X=x')=z\}} - 1_{\{Z(X=x)=z\}} | \mathbf{W}] = 0 \quad (17)$$

The assumptions in (15) and (16) can be justified with traditional ignorability assumptions, and hence (15) will hold if X and Z are “as if jointly randomly assigned” within \mathbf{W} , and (16) will hold if X is “as if randomly assigned” within \mathbf{W} . Unfortunately, (17) depends on the covariance between

³I focus here on the identification criteria for $E[Y(X = x, Z = Z(X = x')) - Y(X = x, Z = Z(X = x))]$. The other mechanism specific effects can be treated analogously.

two counterfactual quantities, and therefore, it cannot be identified by ignorability assumptions or even traditional randomized designs over the variables X , Z , and Y .

Since mechanism specific effects cannot be identified with ignorability assumptions over these variables, the analyst should consider the following two questions. First, is it necessary to point identify the mechanism specific effects in order to answer the substantive question of interest. Second, are there (non-ignorability) assumptions that can be utilized to achieve the required level of identification. In the next two sections, I address these questions.

3.2 Interval Identification (Bounding) of Average Mechanism Specific Effects

Given the difficulties of point identifying mechanism specific effects, the analyst may consider the alternative of interval identification. This approach has produced a number of interesting results in the study of average treatment effects (Manski, 1990; Balke and Pearl, 1997; Chickering and Pearl, 1997; Manski, 2003; Quinn, 2008). In this section, I investigate the use of this approach for the interval identification of mechanism specific effects when X , Y , and Z are binary. (I also suppress the conditioning set \mathbf{W} in order to simplify notation.)

Using the tables in Sections 2.1 and 2.2, it is straightforward to bound average mechanism specific effects. If a Sample Average Mechanism Specific Effect (SAMSE) is required, we need only utilize the observed crosstabulation of X , Y and Z , averaging over the minimum values in the sets for the lower bounds and over the maximum values in the sets for the upper bounds.⁴ Table 8 shows an example of this for the effect in (4). Note that it is easy to generate bounds for subsamples as well, and in this table, bounds have been generated for those observations with $X = 1$. I refer to this effect as the Sample Average Mechanism Specific Effect on the Treated (SAMST). Monotonicity assumptions can be included by using Tables 6, 7 (or similar tables for other assumptions).

We may also be able to reduce these bounds by implementing ignorability assumptions. With

⁴If a population average mechanism specific effect is required (and we assume that the sample was independently and identically drawn), we need only take these same bounds, and then account for the sampling variability in the observed crosstabulations.

Table 8: Bounds for the Sample Average Mechanism Specific Effect (SAMSE) and Sample Average Mechanism Specific Effect on the Treated (SAMST) associated with the individual mechanism specific effect in (4) when X , Z , and Y are binary. The notation n_{xyz} refers to the number of observations that take on the specified values of the observed variables.

X	Z	Y	$Y(X = 1, Z(X = 1)) - Y(X = 1, Z(X = 0))$	Min	Max
0	0	0	$\{-1, 0, 1\}$	$n_{000} \cdot -1$	$n_{000} \cdot 1$
0	0	1	$\{-1, 0, 1\}$	$n_{001} \cdot -1$	$n_{001} \cdot 1$
0	1	0	$\{-1, 0, 1\}$	$n_{010} \cdot -1$	$n_{010} \cdot 1$
0	1	1	$\{-1, 0, 1\}$	$n_{011} \cdot -1$	$n_{011} \cdot 1$
1	0	0	$\{-1, 0\}$	$n_{100} \cdot -1$	$n_{100} \cdot 0$
1	0	1	$\{0, 1\}$	$n_{101} \cdot 0$	$n_{101} \cdot 1$
1	1	0	$\{-1, 0\}$	$n_{110} \cdot -1$	$n_{110} \cdot 0$
1	1	1	$\{0, 1\}$	$n_{111} \cdot 0$	$n_{111} \cdot 1$
			SAMSE	$-1 \cdot \frac{n - n_{101} - n_{111}}{n_{100} + n_{110}}$	$1 \cdot \frac{n - n_{100} - n_{110}}{n_{100} + n_{101} + n_{110} + n_{111}}$
			SAMST	$-1 \cdot \frac{n_{100} + n_{110}}{n_{100} + n_{101} + n_{110} + n_{111}}$	$1 \cdot \frac{n_{101} + n_{111}}{n_{100} + n_{101} + n_{110} + n_{111}}$

binary X , Y , and Z , and within strata defined by \mathbf{W} , the population version of the effect in Table 8 can be written as the following (see Appendix B for details):

$$\begin{aligned}
E[Y(X = 1, Z(X = 1)) - Y(X = 1, Z(X = 0)) | \mathbf{W}] = & \\
& E[Y(X = 1, Z = 1) | \mathbf{W}] \{E[Z(X = 1) | \mathbf{W}] - E[Z(X = 0) | \mathbf{W}]\} \\
& + E[Y(X = 1, Z = 0) | \mathbf{W}] \{E[Z(X = 0) | \mathbf{W}] - E[Z(X = 1) | \mathbf{W}]\} \\
& + Cov[Y(X = 1, Z = 1) - Y(X = 1, Z = 0), Z(X = 1) - Z(X = 0) | \mathbf{W}].
\end{aligned}$$

If ignorability of type (15) holds, then $E[Y(X = x, Z = z) | \mathbf{W}] = E[Y | X = x, Z = z, \mathbf{W}]$ and we can estimate $E[Y | X = 1, Z = 1, \mathbf{W}]$ and $E[Y | X = 1, Z = 0, \mathbf{W}]$ with regression (possibly nonparametric). If ignorability of type (16) holds, then $E[Z(X = x) | \mathbf{W}] = E[Z | X = x, \mathbf{W}]$ and we can again estimate $E[Z | X = 0, \mathbf{W}]$ and $E[Z | X = 1, \mathbf{W}]$ with regression. However, the quantity $Cov[Y(X = 1, Z = 1) - Y(X = 1, Z = 0), Z(X = 1) - Z(X = 0)]$ cannot be estimated, and must be bounded or assumed to be negligible.

With binary X , Y , and Z , bounding of the covariance can be accomplished by noting that the expected values of $Y(X = 1, Z = 1) - Y(X = 1, Z = 0)$ and $Z(X = 1) - Z(X = 0)$ may be estimated due to the ignorability assumptions of (15) and (16), and the variances of both quantities can be bounded because they only take the values -1 , 0 , and 1 . Therefore $Cov[Y(X = 1, Z =$

1) $-Y(X = 1, Z = 0), Z(X = 1) - Z(X = 0)]$ can be bounded by the following:

$$LB = -1 \cdot \max\{\sqrt{V[Y(X = 1, Z = 1) - Y(X = 1, Z = 0)]}\} \cdot \max\{\sqrt{V[Z(X = 1) - Z(X = 0)]}\}$$

$$UB = 1 \cdot \max\{\sqrt{V[Y(X = 1, Z = 1) - Y(X = 1, Z = 0)]}\} \cdot \max\{\sqrt{V[Z(X = 1) - Z(X = 0)]}\}$$

In many applications, monotonicity or ignorability assumptions by themselves will not provide bounds that are sufficiently informative, and the analyst may choose to utilize additional assumptions. In the next section, I discuss assumptions that will provide point identification of mechanism specific effects.

3.3 Point Identifying Mechanism Specific Effects

Equation (14) implies that the mechanism specific effects may be identified if the ignorability conditions (15) and (16) hold, and if the non-ignorability condition (17) holds. However, given the lack of experimental justification for (17) and the associated lack of intuition, the analyst may prefer alternative point identification criteria.

3.3.1 The Special Case of Linear Structural Equation Models

In Linear Structural Equation Models (SEMs), the definition and identification of mechanism specific effects is straightforward. For example, with a binary treatment variable X , a continuous intermediate variable Z , and a continuous outcome variables Y , we might use the following structural equation model:

$$Z = \gamma_0 + \gamma_1 X + U_Z$$

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + U_Y$$

where the errors $\{U_Z, U_Y\}$ are assumed to be random noise. In this model,

$$\begin{aligned}
E[Z|X] &= \gamma_0 + \gamma_1 X \\
E[Y|X] &= \beta_0 + \beta_1 X + \beta_2 E[Z|X] \\
&= \beta_0 + \beta_1 X + \beta_2(\gamma_0 + \gamma_1 X) \\
&= (\beta_0 + \beta_2 \gamma_0) + (\beta_1 + \beta_2 \gamma_1) X,
\end{aligned}$$

where γ_1 is often interpreted as the effect of X on Z , β_2 as the effect of Z on Y and β_1 is often interpreted as the direct effect of X on Y . For this model, the total effect of X on Y can be reconstructed from these effects as $\beta_1 + \beta_2 \gamma_1$, where $\beta_2 \gamma_1$ is often interpreted as an indirect (or path specific) effect (Haavelmo, 1943; Simon, 1953; Goldberger, 1972; Duncan, 1985). Notice that with the constant causal effects implicit in this model, we can write the total individual causal effects in terms of potential variables as the following:

$$\begin{aligned}
Y_i(X_i = 1) - Y_i(X_i = 0) &= (\beta_0 + \beta_1 \cdot 1 + \beta_2 Z_i(X_i = 1) + U_{Zi}) - (\beta_0 + \beta_1 \cdot 0 + \beta_2 Z_i(X_i = 0) + U_{Zi}) \\
&= \beta_1 + \beta_2 (Z_i(X_i = 1) - Z_i(X_i = 0)) \\
&= \beta_1 + \beta_2 \{ \gamma_0 + \gamma_1 \cdot 1 + U_{Zi} - (\gamma_0 + \gamma_1 \cdot 0 + U_{Zi}) \} \\
&= \beta_1 + \beta_2 \cdot \gamma_1
\end{aligned}$$

and therefore, the classical notion of the total effect in the linear SEM corresponds to the counterfactual definition from Section 2.1. Furthermore, we can decompose this effect into mechanism specific effects as in Definition 6.

$$\begin{aligned}
Y_i(X_i = 1, Z_i(X_i = 0)) - Y_i(X_i = 0, Z_i(X_i = 0)) &= (\beta_0 + \beta_1 \cdot 1 + \beta_2 Z_i(X_i = 0) + U_{Zi}) \\
&\quad - (\beta_0 + \beta_1 \cdot 0 + \beta_2 Z_i(X_i = 0) + U_{Zi}) \\
&= \beta_1 \\
&= Y_i(X_i = 1, Z_i(X_i = 1)) - Y_i(X_i = 0, Z_i(X_i = 1)) \\
&= Y_i(X_i = 1, z) - Y_i(X_i = 0, z)
\end{aligned}$$

Therefore, for this model, the “direct effect” β_1 is equivalent to the “all other mechanisms” specific effect for both $Z_i(X_i = 0)$ and $Z_i(X_i = 1)$, and is equivalent to the controlled direct effect for any realized value of $Z_i = z$. This explains why mechanism specific effects can be easily derived in the linear SEM, and it also demonstrates that the homogenous causal effects assumption trivializes the identification condition in (17). The “indirect effect” in this model can be similarly derived:

$$\begin{aligned}
Y_i(X_i = 1, Z_i(X_i = 1)) - Y_i(X_i = 1, Z_i(X_i = 0)) &= (\beta_0 + \beta_1 \cdot 1 + \beta_2 Z_i(X_i = 1) + U_{Zi}) \\
&\quad - (\beta_0 + \beta_1 \cdot 1 + \beta_2 Z_i(X_i = 0) + U_{Zi}) \\
&= \beta_2(Z_i(X_i = 1) - Z_i(X_i = 0)) \\
&= \beta_2(\gamma_0 + \gamma_1 \cdot 1 + U_{Zi}) - \beta_2(\gamma_0 + \gamma_1 \cdot 0 + U_{Zi}) \\
&= \beta_2 \cdot \gamma_1 \\
&= Y_i(X_i = 0, Z_i(X_i = 1)) - Y_i(X_i = 0, Z_i(X_i = 0)).
\end{aligned}$$

3.3.2 Combining Ignorability with the Assumption of no Interaction

While the assumptions of the linear SEM will often seem implausible, there is a weaker assumption that in combination with the ignorability assumptions (13) and (15), will provide point identification for mechanism specific effects (Robins, 2003). When it is plausible to assume no interactive effect between X and Z on Y , then as with linear SEMs the processes of defining and identifying mechanism specific effects becomes greatly simplified. Specifically, we assume that the average controlled direct effect of X on Y does not depend on the controlling value of Z .

$$E[Y(X = x', Z = z) - Y(X = x, Z = z)] = E[Y(X = x', Z = z') - Y(X = x, Z = z')] \quad (18)$$

We should notice two things about this assumption. First, under (18) the equivalence of the average controlled direct effects implies that all average “all other mechanisms” direct effects will also be equivalent to all controlled direct effects,

$$E[Y(X = x', Z = z) - Y(X = x, Z = z)] = E[Y(X = x', Z = Z(X = x)) - Y(X = x, Z = Z(X = x))],$$

and that the indirect effects can be calculated by taking the difference between the total effects and the controlled direct effect for any value of z .

$$\begin{aligned}
E[Y(X = x') - Y(X = x)] &= E[Y(X = x', Z = z) - Y(X = x, Z = z)] \\
&\quad + E[Y(X = x, Z = Z(X = x')) - Y(X = x, Z = Z(X = x))] \\
&= E[Y(X = x', Z = z) - Y(X = x, Z = z)] \\
&\quad + E[Y(X = x', Z = Z(X = x')) - Y(X = x', Z = Z(X = x))]
\end{aligned}$$

Second, both the total effect $E[Y(X = x') - Y(X = x)]$ and the controlled direct effect $E[Y(X = x', Z = z) - Y(X = x, Z = z)]$ can be identified with traditional ignorability assumptions. Therefore, if (13) and (15) hold, then we can identify total effects, controlled direct effects, and mechanism specific effects, we can estimate these effects non-parametrically, and we can potentially diagnose violations of the “no interaction” assumption (18).

In the next section, I present a number of illustrative applications in order to motivate the use of mechanism specific effects. Using examples from American and Comparative politics, I demonstrate nonparametric and semiparametric estimation as well as bounding approaches for mechanism specific effects.

4 Illustrative Applications

4.1 Sex Bias in Graduate Admissions at UC Berkeley

In a classic case of Simpson’s Paradox, admissions data from University of California at Berkeley in 1973 showed clear prima facie evidence of sex bias in graduate admissions. For the six largest majors,⁵ female applicants were accepted at a 14% lower rate than male applicants. However, Bickel et al. (1975) showed that when this analysis was stratified at the department level and pooled correctly, the advantage was reversed. If we are interested in the effects of gender (instead of the effects of perceived gender), how can one interpret these results causally when the conditioning variable (department of application) is potentially affected by the primary treatment variable (sex)?

⁵For this illustrative example, the data were taken from Freedman et al. (1991)

The definitions and methods from this paper provide an approach to interpretation. If we treat Sex as the primary treatment variable (X), department of application as the concomitant variable Z (where for simplicity in presentation I have coded departments into hard (1) and easy (0)), and admission as the outcome variable (Y). Then we can calculate prima facie⁶ mechanism specific estimates for the gender effect on admission that goes through department choice, and the gender effect on admission that does not go through department choice (see Appendix C for details).

Figure 3 shows these average results (with 95% confidence intervals generated by 1000 bootstrap replications). Notice that while the prima facie average total female effect on admission is negative, the portion of this effect that goes through department choice is highly negative regardless of whether gender is held as female or male. Therefore, because mechanism specific effects must add up to the total effect, the average “direct” effect of being female has a positive effect on admission chances regardless of whether department choice is held at “chosen as if male” or “chosen as if female” (note however that the “direct” female effect when department choice is held at “chosen as if female” cannot be distinguished from zero at the 95% confidence level).

4.2 The Effect of Ethnic Heterogeneity on the Probability of Civil War Onset

In an influential paper on the determinants of the onset of civil war, Fearon and Laitin (2003) presents evidence that ethnic heterogeneity may not have the explanatory role that is typically assumed by the academic, policy and media communities.

... after controlling for per capita income, more ethnically or religiously diverse countries have been no more likely to experience significant civil violence in this period... The factors that explain which countries have been at risk for civil war are not their ethnic or religious characteristics but rather the conditions that favor insurgency. (Fearon and Laitin, 2003, p.75)

However, the causal interpretation of these results hinges critically on how we interpret “after controlling for per capita income.” In typical usage, we control for a variable in order to remove a spurious cause, but in this case, country level ethnic heterogeneity is measured prior to the first measurement of per capita income for each country, and ethnic heterogeneity is constant for each

⁶I did not condition on any variables for this analysis and all estimates should be considered only prima facie evidence.

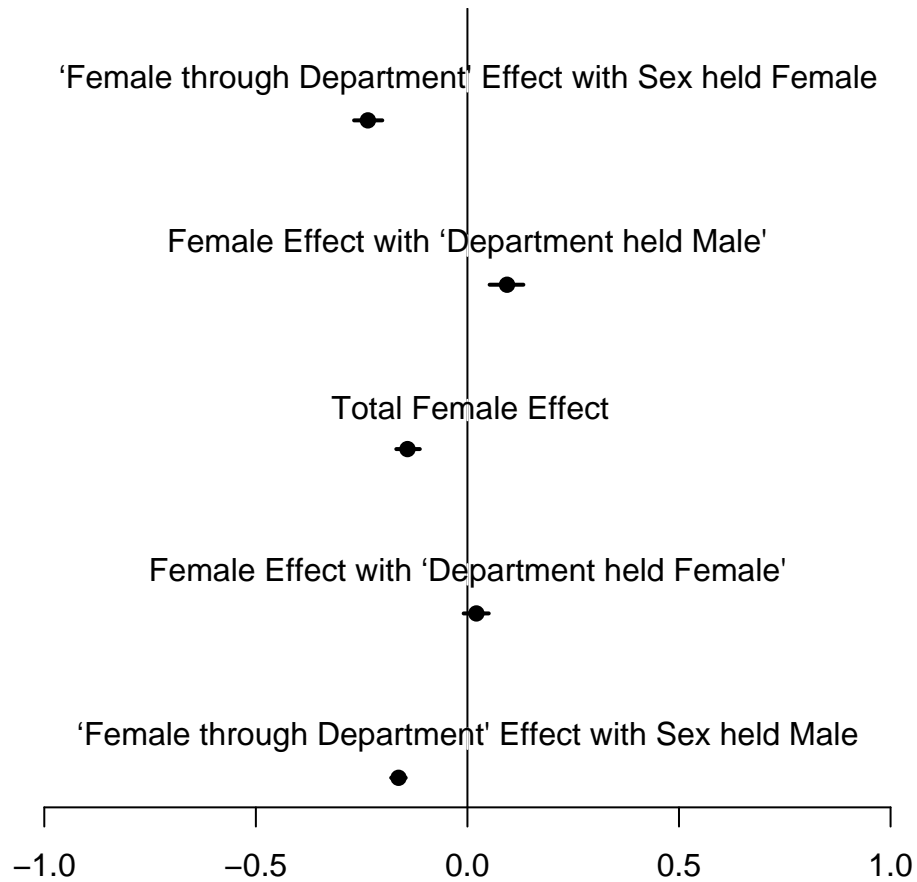


Figure 3: *Prima facie estimates of population average mechanism specific effects of Sex on the probability of Admission with 95% confidence intervals (1000 bootstrap replications). The effect “ ‘Female through Department’ Effect with Sex held Female” is short for “the effect of changing the department choice from as if male to as if female while holding sex as female”. The effect “Female Effect with ‘Department held Male’ ” is short for “the effect of changing from male to female with the department choice held as if male”. The other mechanism specific effects can be stated analogously.*

country throughout the data set. Therefore, it seems more proper to think of per capita income as “post treatment” to ethnic heterogeneity in the sense of King (1991). The methods introduced in this paper allow the inclusion of “post treatment” variables in the model and provide an interpretation in terms of mechanism specific effects. For this analysis, I treat ethnic fractionalization as the primary treatment variable (X), I treat GDP per capita as the concomitant variable (Z), and I treat civil war onset as the outcome variable (Y).

Figure 4 presents the two estimated decompositions of the average total ethnic heterogeneity effect into income mechanism specific effects and “remaining mechanisms” specific effects (the minimum value of ethnic fractionalization was used as the baseline treatment value). In both plots, the top of the curve represents the total average effect of changing ethnic heterogeneity from its minimum value to the value on the x -axis. In Figure 4 (a), the red vertical distance represents the income mechanism specific effect, where ethnicity is held at the value on the x -axis ($E[Y(X = x, Z(X = x)) - Y(X = x, Z(X = x_{min}))]$). The blue vertical distance represents the ‘remaining mechanisms’ specific effect, where income is held at level predicted by minimum ethnic heterogeneity for each observation ($E[Y(X = x, Z(X = x_{min})) - Y(X = x_{min}, Z(X = x_{min}))]$). In Figure 4 (b), the red vertical distance represents the income mechanism specific effect, where ethnicity is held at the minimum value ($E[Y(X = x_{min}, Z(X = x)) - Y(X = x_{min}, Z(X = x_{min}))]$). The blue vertical distance represents the ‘remaining mechanisms’ specific effect, where income is held at level predicted by the x -axis value of ethnic heterogeneity for each observation ($E[Y(X = x, Z(X = x)) - Y(X = x_{min}, Z(X = x))]$). See appendix D for details. Note that for both plots, the control variables from Fearon and Laitin (2003) Table 1 column 1 are used. However, because there is no clear causal order over many of these variables, the “total” effects and the decomposition in this plot are more properly interpreted as averages over controlled direct effects. Note as well that this analysis utilizes a great deal of extrapolation (see King and Zeng (2006) for a discussion on the dangers of extrapolation).

Figure 4 shows that if ethnic heterogeneity is causally prior to GDP (and all the other assumptions specified in Section 3 hold), then we can interpret the results from the model in Fearon and Laitin (2003) in terms of mechanism specific effects. To be precise, ethnic heterogeneity appears to *have an effect* on the probability of civil war onset, but the majority of this effect goes through the income mechanism.

4.3 The Effect of Oil on the Probability of Civil War Onset Due to the Weakening of State Capacity

A number of recent influential papers have noted the connection between natural resources and civil war. Fearon and Laitin (2003) and Collier and Hoeffler (2004) are two prominent examples, and Ross (2004) provides a comprehensive summary of work in this area, concluding the following:

... a close look at both the quantitative and qualitative studies suggests four regularities... The first pattern is that oil exports are linked to the onset of conflict... What causal mechanisms produce these outcomes? Several studies have emphasized that we still know little about the processes that tie natural resources to conflict. (Ross, 2004, p. 338)

In this illustrative application, I investigate the so-called weak states mechanism as outlined in the quote from Fearon and Laitin (2003) on the first page of this manuscript. Using their data on each country's first year, I treat their binary measure of oil production as the primary treatment variable (X), I utilize a binary measure of state weakness as the concomitant variable (Z), and their indicator of civil war onset is used as the outcome variable (Y). In this illustrative application, I have employed a binary measure of state weakness utilized in the Humphreys (2005) analysis of the weak states mechanism: whether a country was either unstable or "anocratic" as coded in Fearon and Laitin (2003). Figure 5 shows the results from a bounding analysis on the Sample Average Mechanism Specific Effects on the Treated (SAMST). For this sample, I have reported the logical bounds on the average mechanism specific effects for the oil producing countries because this corresponds closely to a policy question of interest: "What would have happened to the oil producing countries if oil production had been conducted in such a way so as not to affect state apparatuses?" Notice that the mechanism specific effect of oil production on the probability of civil

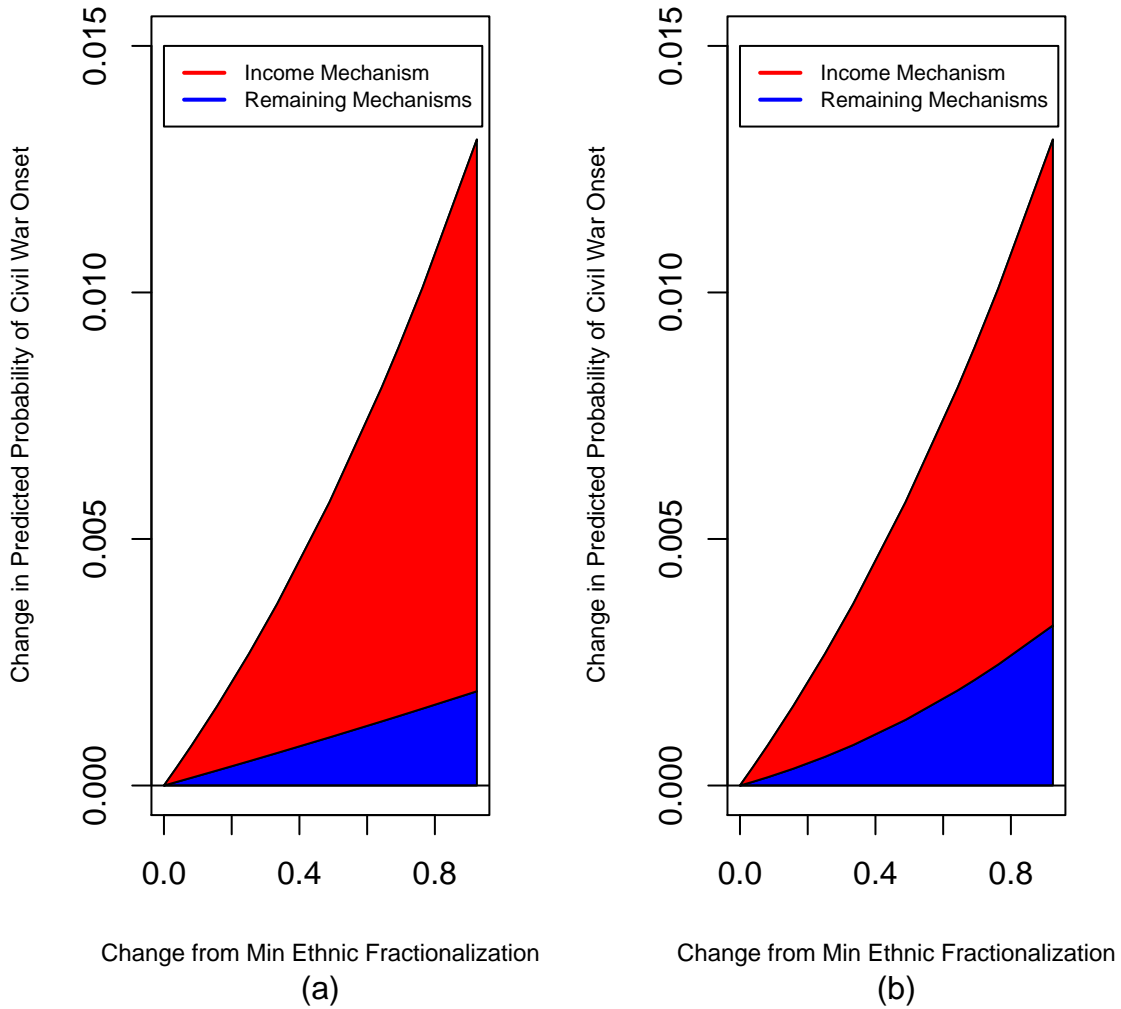


Figure 4: *The two estimated decompositions of the “average total ethnic heterogeneity effect on the probability of civil war onset” into the income mechanism specific effects and the “remaining mechanisms” specific effects, with the minimum value of ethnic fractionalization used as the baseline treatment value. (a) The income mechanism specific effect is $E[Y(X = x, Z(X = x)) - Y(X = x, Z(X = x_{min}))]$. The ‘remaining mechanisms’ specific effect is $E[Y(X = x, Z(X = x_{min})) - Y(X = x_{min}, Z(X = x_{min}))]$. (b) The income mechanism specific effect is $E[Y(X = x_{min}, Z(X = x)) - Y(X = x_{min}, Z(X = x_{min}))]$. The ‘remaining mechanisms’ specific effect is $E[Y(X = x, Z(X = x)) - Y(X = x_{min}, Z(X = x))]$. For both plots, the control variables from Fearon and Laitin (2003) Table 1 column 1 are used. Also note that because there is no clear causal order over many of these control variables, so the “total” effects and the decomposition in this plot are more properly interpreted as averages over controlled direct effects.*

war that is due to state weakness addresses this issue (see SAMST in Table 8 for a description of the bounds for this effect).

The top line in Figure 5 (a) shows the the logical bounds for this effect and we see that this mechanism specific effect can be at most slightly positive. However, if we include the seemingly reasonable monotonicity assumption that state weakness has a non-negative effect on the probability of civil war onset when oil production status is held constant (Figure 5 (b)), then *the bounds show the mechanism specific effect to be non-positive*. This result seems to run counter to the standard logic that significant oil production tends to weaken state apparatuses, and that weak state apparatuses increase the chances of civil war onset. However, of the 18 countries that were designated as oil producers for their first year in the study,⁷ only two experienced civil wars, and these were not designated as weak states under the discussed measure. Furthermore, we might expect the indirect effect of oil through state capacity to be delayed. Figure 6, shows that we cannot rule out this scenario because the bounds on the indirect effect include positive values once we change the dependent variable to indicate whether a civil war happened in the first year or the second year. This plot further shows the changes in the bounds as we change the dependent variable to indicate whether a civil war happened in the first 2 - 15 years.

5 Conclusion

In this paper, I have shown that the use of counterfactuals in the formal definition of causal mechanism specific effects provides a number of benefits. First, counterfactuals clarify the definition of causal mechanisms and illuminate the policy implications of mechanism specific effects. Second, counterfactuals facilitate the statement of conditions for point or interval identification of average mechanism specific effects— *even when causal effect heterogeneity is allowed*. Third, the decomposition of total effects into mechanism specific effects at the individual level allows us to specify and answer questions that may not have been apparent to us without this technology. Finally, this paper

⁷All these countries were stable as oil producers in that they were designated as oil producers throughout at least their first decade in the data set.

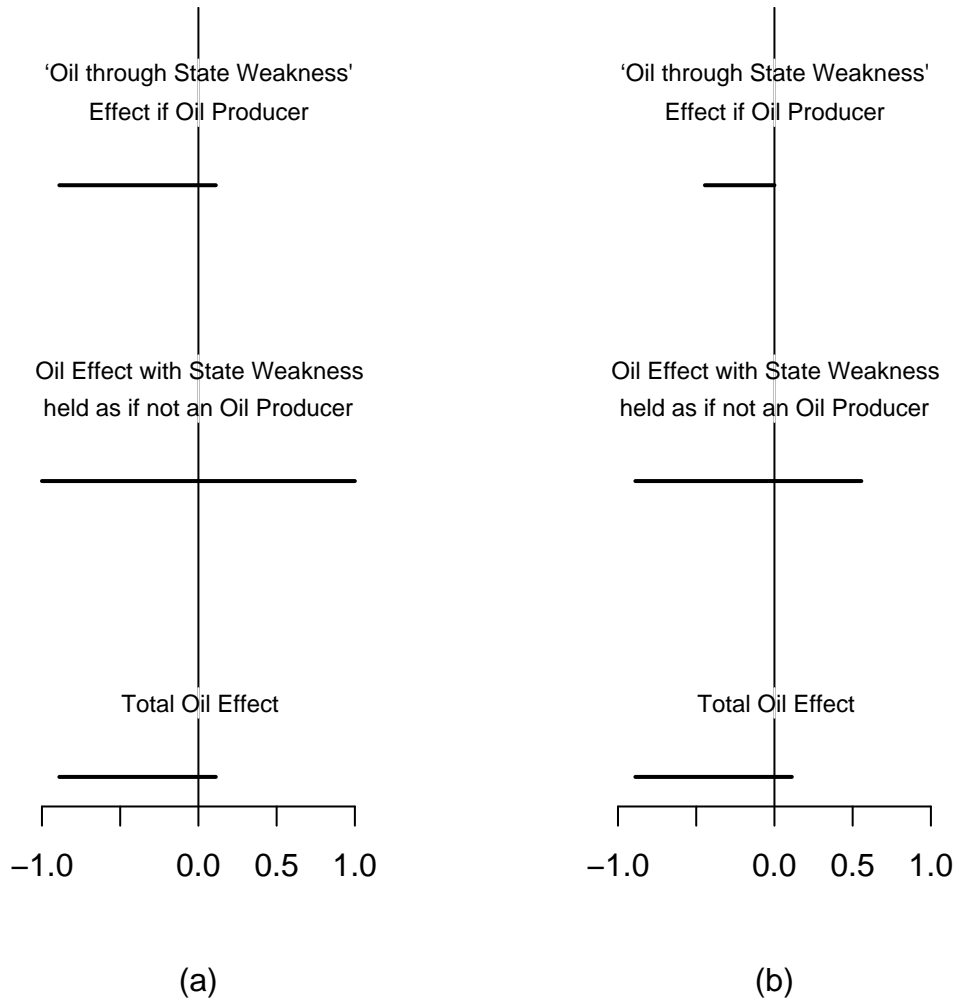


Figure 5: *Logical bounds for the sample average mechanism specific effects of oil production on the probability of civil war onset for oil producing nations in the first year for each country in the data set (see SAMST in Table 8 for details on panel (a)). (a) No auxiliary assumptions made (b) State weakness is assumed to have a non-negative effect on the probability of civil war onset when oil production status is held constant.*

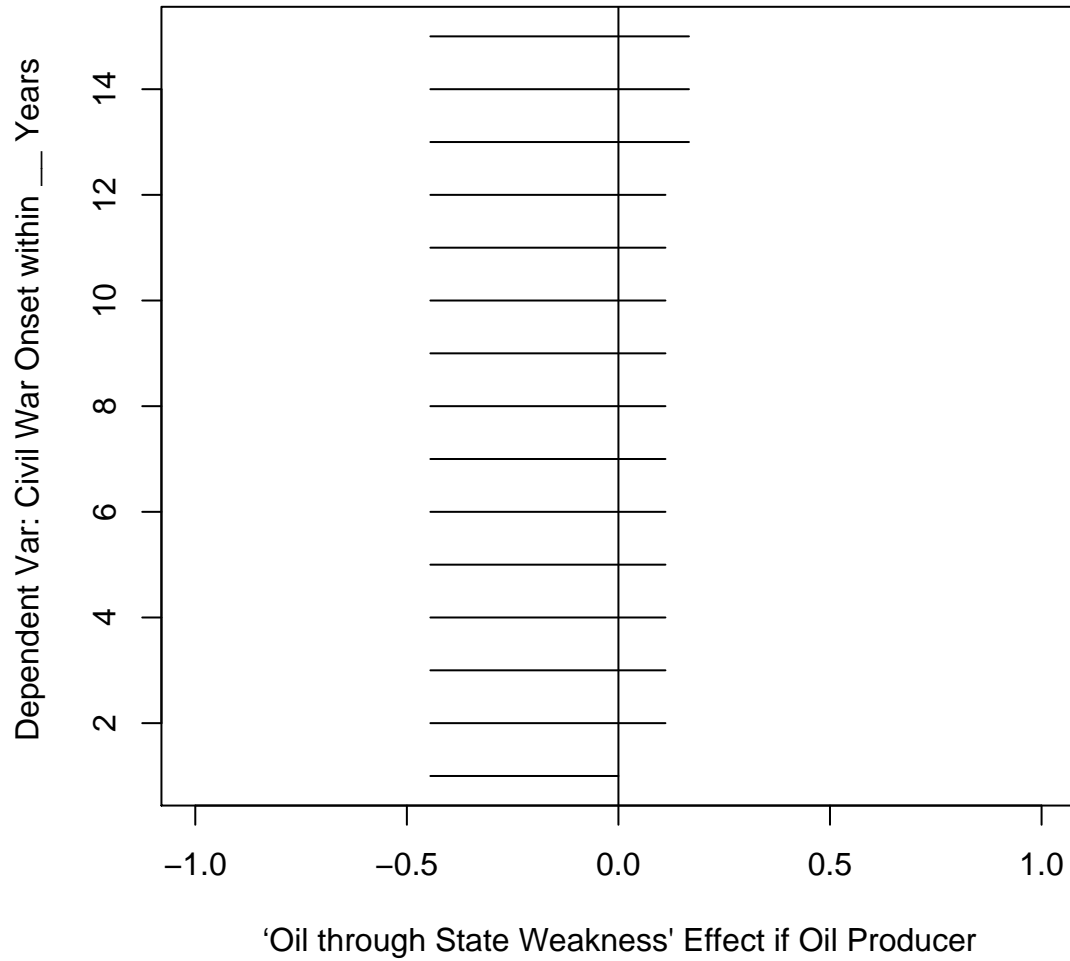


Figure 6: *Among first year oil producing nations, logical bounds for the sample average ‘weak states’ mechanism specific effect on the probability of civil war onset within up to the first fifteen years of state inclusion in the data set. State weakness is assumed to have a non-negative effect on the probability of civil war onset when oil production status is held constant.*

has shown that learning about causal mechanisms is *more difficult* than is typically understood. Therefore, political scientists should perhaps re-evaluate their objectives when making inference about causal mechanisms (i.e. is point identification plausible or even necessary). Furthermore, in order to make reasonable inference about causal mechanisms, we may need to borrow strength from a number of different types of causal assumptions (ignorability will often be insufficient). In future work, I will utilize a Bayesian approach to combine different types of causal assumptions and to weaken the assumptions made in this paper.

Appendix A: Derivation of the Identification Formula for Mechanism Specific Effects

$$\begin{aligned} E[Y(X = x, Z = Z(X = x')) - Y(X = x, Z = Z(X = x))] \\ = E[E[Y(X = x, Z = Z(X = x')) - Y(X = x, Z = Z(X = x)) | \mathbf{W}]] \end{aligned}$$

$$\begin{aligned} E[Y(X = x, Z = Z(X = x')) - Y(X = x, Z = Z(X = x)) | \mathbf{W}] &= \\ = E\left[\sum_z \{Y(X = x, Z = z)1_{\{Z(X=x')=z\}} - Y(X = x, Z = z)1_{\{Z(X=x)=z\}}\} | \mathbf{W}\right] \\ = \sum_z E[Y(X = x, Z = z)1_{\{Z(X=x')=z\}} - Y(X = x, Z = z)1_{\{Z(X=x)=z\}} | \mathbf{W}] \\ = \sum_z \{E[Y(X = x, Z = z)1_{\{Z(X=x')=z\}} | \mathbf{W}] - E[Y(X = x, Z = z)1_{\{Z(X=x)=z\}} | \mathbf{W}]\} \\ = \sum_z \{E[Y(X = x, Z = z) | \mathbf{W}]E[1_{\{Z(X=x')=z\}} | \mathbf{W}] + Cov[Y(X = x, Z = z), 1_{\{Z(X=x')=z\}} | \mathbf{W}] \\ - E[Y(X = x, Z = z) | \mathbf{W}]E[1_{\{Z(X=x)=z\}} | \mathbf{W}] - Cov[Y(X = x, Z = z), 1_{\{Z(X=x)=z\}} | \mathbf{W}]\} \\ = \sum_z \{E[Y(X = x, Z = z) | \mathbf{W}]E[1_{\{Z(X=x')=z\}} | \mathbf{W}] \\ - E[Y(X = x, Z = z) | \mathbf{W}]E[1_{\{Z(X=x)=z\}} | \mathbf{W}] \\ + Cov[Y(X = x, Z = z), 1_{\{Z(X=x')=z\}} - 1_{\{Z(X=x)=z\}} | \mathbf{W}]\} \end{aligned}$$

Appendix B: Derivation of the Identification Formula for Mechanism Specific Effects

$$\begin{aligned}
& E[Y(1, Z(X = 1)) - Y(1, Z(X = 0)) | \mathbf{W}] = \\
& E[Y(X = 1, Z = 1)Z(X = 1) + Y(X = 1, Z = 0) \cdot (1 - Z(X = 1)) | \mathbf{W}] \\
& - E[Y(X = 1, Z = 1)Z(X = 0) + Y(X = 1, Z = 0) \cdot (1 - Z(X = 0)) | \mathbf{W}] \\
& = E[Y(X = 1, Z = 1) | \mathbf{W}]E[Z(X = 1) | \mathbf{W}] + Cov[Y(X = 1, Z = 1), Z(X = 1) | \mathbf{W}] \\
& + E[Y(X = 1, Z = 0) | \mathbf{W}]E[(1 - Z(X = 1)) | \mathbf{W}] + Cov[Y(X = 1, Z = 0), 1 - Z(X = 1) | \mathbf{W}] \\
& - E[Y(X = 1, Z = 1) | \mathbf{W}]E[Z(X = 0) | \mathbf{W}] - Cov[Y(X = 1, Z = 1), Z(X = 0) | \mathbf{W}] \\
& - E[Y(X = 1, Z = 0) | \mathbf{W}]E[(1 - Z(X = 0)) | \mathbf{W}] - Cov[Y(X = 1, Z = 0), (1 - Z(X = 0)) | \mathbf{W}] \\
& = E[Y(X = 1, Z = 1) | \mathbf{W}] \{E[Z(X = 1) | \mathbf{W}] - E[Z(X = 0) | \mathbf{W}]\} \\
& + E[Y(X = 1, Z = 0) | \mathbf{W}] \{E[1 - Z(X = 1) | \mathbf{W}] - E[1 - Z(X = 0) | \mathbf{W}]\} \\
& + Cov[Y(X = 1, Z = 1), Z(X = 1) | \mathbf{W}] - Cov[Y(X = 1, Z = 0), Z(X = 1) | \mathbf{W}] \\
& - Cov[Y(X = 1, Z = 1), Z(X = 0) | \mathbf{W}] + Cov[Y(X = 1, Z = 0), (Z(X = 0)) | \mathbf{W}] \\
& = E[Y(X = 1, Z = 1) | \mathbf{W}] \{E[Z(X = 1) | \mathbf{W}] - E[Z(X = 0) | \mathbf{W}]\} \\
& + E[Y(X = 1, Z = 0) | \mathbf{W}] \{E[Z(X = 0) | \mathbf{W}] - E[Z(X = 1) | \mathbf{W}]\} \\
& + Cov[Y(X = 1, Z = 1) - Y(X = 1, Z = 0), Z(X = 1) - Z(X = 0) | \mathbf{W}]
\end{aligned}$$

Appendix C: Formulas for the Prima Facie Estimates in Section 4.1

As shown in Definition 7, there are two ways to decompose the total effect, and given that X , Y , and Z are all binary for this example, we can estimate the total effect and the decomposition in mechanism specific effects with the following formulas. The notation p_{xzy} refers to the fraction of observations that take on the specified values of the observed variables, and the subscripts in this

notation can be replaced by + signs to indicate sums over proportions (e.g. $p_{1+1} = p_{101} + p_{111}$).

$$\begin{aligned}\widehat{E}[Y(X = 1) - Y(X = 0)]_{pf} &= \frac{p_{1+1}}{p_{1++}} - \frac{p_{0+1}}{p_{0++}} \\ \widehat{E}[Y(X = 1, Z(X = 1)) - Y(X = 1, Z(X = 0))]_{pf} &= \frac{p_{111}}{p_{11+}} \cdot \left(\frac{p_{11+}}{p_{1++}} - \frac{p_{01+}}{p_{0++}} \right) \\ &\quad + \frac{p_{101}}{p_{10+}} \cdot \left(\frac{p_{01+}}{p_{0++}} - \frac{p_{11+}}{p_{1++}} \right) \\ \widehat{E}[Y(X = 0, Z(X = 1)) - Y(X = 0, Z(X = 0))]_{pf} &= \frac{p_{011}}{p_{01+}} \cdot \left(\frac{p_{11+}}{p_{1++}} - \frac{p_{01+}}{p_{0++}} \right) \\ &\quad + \frac{p_{001}}{p_{00+}} \cdot \left(\frac{p_{01+}}{p_{0++}} - \frac{p_{11+}}{p_{1++}} \right)\end{aligned}$$

Appendix D: Estimation Process for Section 4.2

The following process was used to generate the size of the total and mechanism specific effects for each value x of ethnic heterogeneity in Figure 4. Note that while logistic regression was used in this case, nonparametric regression could potentially be substituted.

Process to calculate the “total” ethnic heterogeneity effect:

1. Fit the logistic regression in Fearon and Laitin (2003) Table 1 column 1 but with income (GDP per capita) removed from the model.
2. Using the results from the regression with income removed, predict the probability of civil war onset for every observation in the data set with ethnic heterogeneity set at x_{min} (keep the other control variables at their realized values for each observation).
3. Repeat the process in Step 2 with ethnic heterogeneity set to x and take the difference in these predicted probabilities for each observation of moving from x_{min} to x .
4. Average over these observation specific differences to get the total ethnicity effect of moving from x_{min} to x .

Process to calculate the “remaining mechanisms” ethnic heterogeneity effect in Figure 4 (a):

1. Fit the logistic regression in Fearon and Laitin (2003) Table 1 column 1.
2. Create a discrete version of GDP. (For this plot GDP was cut into ten bins based on its deciles.)
3. Run a multinomial regression of discrete GDP on the remaining variables.
4. Using the results from the logistic regression, with ethnic heterogeneity set at x_{min} and GDP set at the midpoint of the first discrete bin, calculate the predicted probability of onset for each observation in the data set (keep the other control variables at their realized values for each observation).

5. Repeat Step 4 with ethnic heterogeneity set at x , and take the difference in predicted probabilities for each observation.
6. Using the results from the multinomial regression, with ethnic heterogeneity set at x_{min} , calculate the predicted probability of GDP falling in the first bin for each observation in the data set (keep the other control variables at their realized values for each observation). For each observation, multiply this predicted probability by the difference in predicted probabilities from Step 5.
7. Repeat Steps 4, 5 and 6 for each discrete value of GDP. For each observation in the data set, sum the results from Step 6 over the discrete values of GDP.
8. Average over the observation specific quantities in Step 7 to get the average “remaining mechanisms” ethnicity effect of moving from x_{min} to x in Figure 4 (a).

Process to calculate the “remaining mechanisms” ethnic heterogeneity effect in Figure 4 (b):

1. Fit the logistic regression in Fearon and Laitin (2003) Table 1 column 1.
2. Create a discrete version of GDP. (For this plot GDP was cut into ten bins based on its deciles.)
3. Run a multinomial regression of discrete GDP on the remaining variables.
4. Using the results from the logistic regression, with ethnic heterogeneity set at x_{min} and GDP set at the midpoint of the first discrete bin, calculate the predicted probability of onset for each observation in the data set (keep the other control variables at their realized values for each observation).
5. Repeat Step 4 with ethnic heterogeneity set at x , and take the difference in predicted probabilities for each observation.
6. Using the results from the multinomial regression, with ethnic heterogeneity set at x , calculate the predicted probability of GDP falling in the first bin for each observation in the data set (keep the other control variables at their realized values for each observation). For each observation, multiply this predicted probability by the difference in predicted probabilities from Step 5.
7. Repeat Steps 4, 5 and 6 for each discrete value of GDP. For each observation in the data set, sum the results from Step 6 over the discrete values of GDP.
8. Average over the observation specific quantities in Step 7 to get the average “remaining mechanisms” ethnicity effect of moving from x_{min} to x in Figure 4 (b).

References

- Balke, A., and J. Pearl. 1997. "Bounds on Treatment Effects from Studies with Imperfect Compliance." *Journal of the American Statistical Association* 92(439).
- Bickel, P.J, EA Hammel, and JW O'Connell. 1975. "Sex Bias in Graduate Admissions: Data from Berkeley." *Science* 187(4175):398–404.
- Campbell, Angus, Philip Converse, Warren E. Miller, and Donald Stokes. 1960. *The American Voter*. New York: John Wiley.
- Chickering, D.M., and J. Pearl. 1997. "A clinicians tool for analyzing non-compliance." *Computing Science and Statistics* 29(2):424–431.
- Collier, David, and Henry E. Brady. 2004. *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. Lanham, MD: Rowman & Littlefield.
- Collier, P., and A. Hoeffler. 2004. "Greed and grievance in civil war."
- Cox, Gary W. 1997. *Making Votes Count: Strategic Coordination in the World's Electoral Systems*. Cambridge: Cambridge University Press.
- Duncan, O.D. 1985. "Path Analysis: Sociological Examples." *Causal Models in the Social Sciences*
- Fearon, James D., and David D. Laitin. 2003. "Ethnicity, Insurgency, and Civil War." *American Political Science Review* 97:75–90.
- Frangakis, C.E., and D.B. Rubin. 2002. "Principal Stratification in Causal Inference." *Biometrics* 58(1):21–29.
- Freedman, D., R. Pisani, R. Purves, and A. Adhikari. 1991. "Statistics (2nd edn)."
- Goldberger, Arthur S. 1972. "Structural Equation Models in the Social Sciences." *Econometrica* 40:979–1001.
- Haavelmo, Trygve. 1943. "The Statistical Implications of of a System of Simultaneous Equations." *Econometrica* 11:1–12.
- Hall, P.A. 2003. "Aligning Ontology and Methodology in Comparative Research." *Comparative Historical Analysis in the Social Sciences* pp. 373–404.
- Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81:945–960.
- Humphreys, M. 2005. "Natural Resources, Conflict, and Conflict Resolution: Uncovering the Mechanisms." *Journal of Conflict Resolution* 49(4):508.
- King, G. 1991. "' Truth' Is Stranger than Prediction, More Questionable than Causal Inference." *American Journal of Political Science* 35(4):1047–53.
- King, G., and L. Zeng. 2006. "The Dangers of Extreme Counterfactuals." *Political Analysis* 14(2):131–159.
- Manski, C. 1990. "Nonparametric Bounds on Treatment Effects." *American Economic Review* 80(2):319–323.

- Manski, C.F. 2003. *Partial Identification of Probability Distributions*. Springer.
- Pearl, J. 2001. "Direct and indirect effects." *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence* pp. 411–420.
- Pearl, Judea. 2000. *Causality: Models, Reasoning, and Inference*. New York: Cambridge University Press.
- Petersen, M.L., S.E. Sinisi, and M.J. van der Laan. 2006. "Estimation of direct causal effects." *Epidemiology* 17(3):276–284.
- Quinn, K.M. 2008. "What Can Be Learned from a Simple Table? Bayesian Inference and Sensitivity Analysis for Causal Effects from 2 x 2 and 2 x 2 x K Tables in the Presence of Unmeasured Confounding." *Working Paper* .
- Robins, J.M. 1986. "A new approach to causal inference in mortality studies with a sustained exposure period-application to control of the healthy worker survivor effect." *Mathematical Modeling* 7:1393–1512.
- Robins, J.M. 2003. "Semantics of causal DAG models and the identification of direct and indirect effects." *Highly Structured Stochastic Systems* pp. 70–81.
- Robins, J.M., and S. Greenland. 1992. "Identifiability and Exchangeability for Direct and Indirect Effects." *Epidemiology* 3(2):143–155.
- Rosato, S. 2003. "The Flawed Logic of Democratic Peace Theory." *American Political Science Review* 97(04):585–602.
- Ross, M.L. 2004. "What Do We Know about Natural Resources and Civil War?" *Journal of Peace Research* 41(3):337.
- Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66(5):688–701.
- Simon, Herbert A. 1953. "Causal Ordering and Identifiability." In *Studies in Econometric Method* (W.C. Hood, and T.C. Hoopmans, editors), New York: John Wiley.