Phantom Counterfactuals

Tara Slough*

March 2, 2022

Contents

*Assistant Professor, New York University. taraslough@nyu.edu

Appendix A Proofs

A1.1 Proof of Proposition 1

Suppose that an experiment manipulates a single treatment Z. Assume:

1. Treatment assignment is ignorable: $Y(Z) \perp Z, Pr(Z = z) \in (0, 1)$.
2. SUTVA: $Y_i(z_i) = Y_i(z_i, \mathbf{z}_{\neg i}) \forall i$.

Consider a dynamic model for which $h^{\emptyset} \neq h^T$. Index sets of non-terminal histories, $h \in H \backslash H^T$ by the cardinality of the set of past actions. In this notation, $h^{\emptyset} \equiv h^0$. The subsequent histories are represented by $h \in H^1$. etc. In this notation, a dynamic model implies that $\exists H^1$.

With this notation, strategy set symmetry, as defined in Definition 1, implies that for any $h \in H^j$, the actor and set of strategies available for all elements H^{j+1} are equivalent, for all $j \in \{0, 1, ..., T-1\}$.

Consider an action, a, in the strategy set of arbitrary node $h \in H$. Denote a variable measuring this action as $\mathcal{A}$. The ATE can be written:

$$\sum_{h \in H^j} Pr(h|Z = 1)E[\mathcal{A}|Z = 1, h = h] - \sum_{h \in H^j} Pr(h|Z = 0)E[\mathcal{A}|Z = 0, h = h] \quad (1)$$

First, consider the first post-treatment action, $j = 0$. Both expectations in Equation 2 are defined. The ATE is both defined and identified given Assumptions 1 and 2 and standard results (i.e. Green and Gerber (2012) Equation 2.3 or Angrist and Pischke (2010) Section 2.2).

Now, consider some $j > 0$. Consider two cases:

1. If a is in the strategy set for all $h \in H^j$, the expression $E[\mathcal{A}|Z = z, h = h]$ is defined. The ATE is both defined and identified.
2. If a is *not* in the strategy set for any $h \in H^j$, the expression $E[\mathcal{A}|Z = z, h = h]$ is undefined for some h. The ATE is undefined, and thus unidentified.

By Definition 1, if a model is strategy set symmetric, it follows from the case of $j = 0$ and Case #1 above that the ATE is identified for all actions. Further, if the model is not strategy set symmetric, it follows from the case of $j = 0$ and Case #2 that the ATE must be identified for at least one outcome (at h^0) and must be unidentified for at least one outcome. ■

A1.2 Proof of Proposition 2

Suppose that an experiment manipulates a single treatment Z. Assume:

1. Treatment assignment is ignorable: $Y(Z) \perp Z, Pr(Z = z) \in (0, 1)$.
2. SUTVA: $Y_i(z_i) = Y_i(z_i, \mathbf{z}_{\neg i}) \forall i$.

Suppose that there exists a strategy-set asymmetric history, denoted $H^s \neq H^T$. Consider an action, a, in the strategy set of arbitrary node $h \in \{H^{S+1}, ..., H^T\}$, where histories are indexed according to the cardinality of past actions. The action a is only in the strategy set at node h if $H^s = S$. Denote a variable measuring

the action S as $\mathcal{S}$ and the variable measuring this action as $\mathcal{A}$.

The SACE of Z on $\mathcal{A}$ is given by:

$$SACE = \sum_{h \in H^S} Pr(\mathcal{S}|Z=1, S(Z=1)=1, S(Z=0)=1) E[\mathcal{A}|Z=1, S(Z=1)=1, S(Z=0)=1] - \sum_{h \in H^S} Pr(\mathcal{S}|Z=0, S(Z=1)=1, S(Z=0)=1) E[\mathcal{A}|Z=0, S(Z=1)=1, S(Z=0)=1] \tag{2}$$

Because $a(Z)$ is defined for for any unit for which $S(Z=1)=1$ and $S(Z=0)=1$, $E[\mathcal{A}|Z=z, S(Z=1)=1, S(Z=0)=1]$ is defined for $Z \in \{0,1\}$. Further, note that the SACE is a conditional ATE in which the stratum defined by $S(Z=1)=1 \cap S(Z=0)=1$. This means that $Pr(\mathcal{S}|Z=z, S(Z=1)=1, S(Z=0)=1)=1 \forall Z \in \{0,1\}$. As such, (2) simplifies to:

$$\begin{aligned} SACE = & E[\mathcal{A}|Z=1, S(Z=1)=1, S(Z=0)=1] - \\ & E[\mathcal{A}|Z=0, S(Z=1)=1, S(Z=0)=1]. \end{aligned} \tag{3}$$

Given that both expectations are defined and standard results showing that conditional ATEs are identified given Assumptions 1 and 2 (i.e., Green and Gerber (2012) Equation 2.3 or Angrist and Pischke (2010) Section 2.2), the SACE is identified. ■

A1.3 Propositions A1-A2 and Proofs

In Propositions A1 and A2, I discuss more generally the relationship between the difference-in-means estimand, Δ, and the $SACE$ in the presence of post-treatment selection. I will use the notation from Table A1, in which $S(Z) \in \{0,1\}$ represents potential outcomes on selection, and subsequent outcome $Y(Z)$ is given by:

$$Y(Z) \in \begin{cases} \mathbb{R} & \text{if } S(Z)=1 \\ * & \text{if } S(Z)=0. \end{cases}$$

When considering this setting, it is important to note that when estimating Δ in the presence of post-treatment selection, researchers have made decisions about how to address the phantom counterfactuals (undefined potential outcomes) in a subsequent outcome $Y(Z)$. These are any units for whom the realization of $Y(Z)=*$. I consider the two modal strategies: (1) conditioning the sample on survival such that all $S(Z)=1$ for all units in the difference-in-means sample; and (2) imputing some constant, $c \in \mathbb{R}$ for $Y(Z)$ whenever $S(Z)=0$. I therefore derive Δ under both strategies.

I use the principal strata defined in Table A1 to express Δ and compare it to $SACE$. To economize notation, I denote membership in a stratum as $\theta \in \{A, T, U, N\}$. With this notation, the SACE can be written as:

$$SACE = E[Y(Z=1)|\theta = A] - E[Y(Z=0)|\theta = A].$$

As in Table A1, the proportion of experimental units in each stratum is π_A, π_T, π_U, and π_N where $\pi_A + \pi_T + \pi_U + \pi_N = 1$.

Proposition A1. *Suppose that there exists post-treatment selection, i.e., $S(Z) = 0$ for $n \geq 1$ units in the experimental sample. Consider the SACE of treatment Z on subsequent outcome $Y(Z)$ and the quantity estimated by a difference-in-means estimator:*

$$\Delta = \overline{Y}(Z = 1) - \overline{Y}(Z = 0).$$

In the absence of additional parametric assumptions, we cannot ascertain the sign or magnitude of the SACE from Δ when:

(i) the sample is conditioned on the realization of $S(Z) = 1$, or

(ii) some constant $c \in \mathbb{R}$ is imputed such that $Y(Z) = c$ for any realization of $S(Z) = 0$.

Proof Consider first the case in which the sample is conditioned on the realization that $S(Z) = 1$. A difference-in-means estimator estimates the quantity:

$$\begin{aligned}\Delta = &\underbrace{\frac{\pi_A}{\pi_A + \pi_T}\overline{Y}(Z = 1|\theta = A) + \frac{\pi_T}{\pi_A + \pi_T}\overline{Y}(Z = 1|\theta = T)}_{\overline{Y}(Z=1|S(Z=1)=1)} - \\ &\underbrace{\left[\frac{\pi_A}{\pi_A + \pi_U}\overline{Y}(Z = 0|\theta = A) + \frac{\pi_U}{\pi_A + \pi_U}\overline{Y}(Z = 0|\theta = U)\right]}_{\overline{Y}(Z=0|S(Z=0)=1)}\end{aligned} \tag{4}$$

Note that $SACE = E[Y(Z = 1)|\theta = A] - E[Y(Z = 0)|\theta = A]$. (4) can be therefore be written:

$$\begin{aligned}\Delta = &\frac{\pi_A^2}{(\pi_A + \pi_T)(\pi_A + \pi_U)}SACE + \frac{\pi_A\pi_U}{(\pi_A + \pi_T)(\pi_A + \pi_U)}\overline{Y}(Z = 1|\theta = A) + \frac{\pi_T}{\pi_A + \pi_T}\overline{Y}(Z = 1|\theta = T) - \\ &\left[\frac{\pi_A\pi_T}{(\pi_A + \pi_T)(\pi_A + \pi_U)}\overline{Y}(Z = 0|\theta = A) + \frac{\pi_U}{\pi_A + \pi_U}\overline{Y}(Z = 0|\theta = U)\right]\end{aligned} \tag{5}$$

Without further parametric assumptions, an estimate of Δ is uninformative about the sign or magnitude of the SACE.

Now consider the case in which a constant $c \in \mathbb{R}$ is imputed for any units for which $S(Z) = 0$. A difference-in-means estimator estimates the quantity:

$$\begin{aligned}\Delta = &\pi_A\overline{Y}(Z = 1|\theta = A) + \pi_T\overline{Y}(Z = 1|\theta = T) + (1 - \pi_A - \pi_T)c - \\ &\left[\pi_A\overline{Y}(Z = 0|\theta = A) + \pi_U\overline{Y}(Z = 0|\theta = U) + (1 - \pi_A - \pi_U)c\right]\end{aligned} \tag{6}$$

Since $SACE = E[Y(Z = 1)|\theta = A] - E[Y(Z = 0)|\theta = A]$. (6) can be therefore be written:

$$\Delta = \pi_A SACE + \pi_T\overline{Y}(Z = 1|\theta = T) - \pi_U\overline{Y}(Z = 0|\theta = U) + c(\pi_T - \pi_U) \tag{7}$$

Without further parametric assumptions, an estimate of Δ is uninformative about the sign or magnitude of the SACE. ■

Proposition A2. *Suppose that there exists post-treatment selection, i.e., $S(Z) = 0$ for $n \geq 1$ units in the experimental sample. Suppose further that selection is exogenous.*

(i) If the sample is conditioned on the realization of $S(Z) = 1$, then $\Delta = SACE$.

(ii) If some constant $c \in \mathbb{R}$ is imputed such that $Y(Z) = c$ for any realization of $S(Z) = 0$, then Δ and SACE maintain the same sign but $|\Delta| < |SACE|$.

Proof of Proposition A2. Within this framework, potential outcomes are fixed. Exogenous post-treatment selection therefore implies that $S(Z = 1) = S(Z = 0)$ for all units in the experimental sample. This implies that the if-treated and if-untreated strata are empty, $\pi_T = \pi_U = 0$. In the case in which the sample is conditioned on $S(Z) = 1$, substituting $\pi_T = \pi_U = 0$ into (5) yields:

$$\Delta = SACE.$$

In the case in which a constant is imputed when $S(Z) = 0$, substituting $\pi_T = \pi_U = 0$ into (7) yields:

$$\Delta = \pi_A SACE.$$

Because $\pi_A \in (0, 1)$, sgn(Δ) = sgn(SACE) and $|\Delta| < |SACE|$. ■

A1.4 Generalization to Average Treatment Effects on the Treated (ATT)

The average treatment effect on the treated (ATT) is a target estimand of many identification-motivated research designs in the social sciences. It is defined as:

$$ATT \equiv E[Y_i(Z = 1)|Z = 1] - E[Y_i(Z = 0)|Z = 1]. \tag{8}$$

Proposition A3 generalizes the main result of Proposition 1 from the ATE to the ATT estimand. Note that there are different research designs under which the identified estimand of interest is the ATT (i.e., difference-in-difference designs and matching). Proposition A3 refers only to the identifying assumptions under a given research design in an effort to maintain generality.

Proposition A3. *Suppose that standard identifying assumptions of the ATT under a research design are satisfied. If a dynamic theory of post-treatment behavior is not strategy set symmetric, then:*

1. *There exists at least one post-treatment behavioral outcome for which the ATT is identified.*
2. *There exists at least one post-treatment behavioral outcome for which the ATT is not identified.*

In a research design in which these standard identifying assumptions hold, if a dynamic theory of post-treatment behavior is strategy set symmetric, then the ATT is identified for all modeled post-treatment behavioral outcomes.

Proof: Refer to the the proof of Proposition 1. Consider the ATT of a treatment, Z, on an action, a, in the strategy set of arbitrary node $h \in H$. Denote a variable measuring this action as $\mathcal{A}$. Given definition of the ATT in (8), the ATT can be written:

$$\sum_{h \in H^j} Pr(h|Z = 1)E[\mathcal{A}(Z = 1)|Z = 1, h = h] - \sum_{h \in H^j} Pr(h|Z = 1)E[\mathcal{A}(Z = 0)|Z = 1, h = h] \tag{9}$$

The remainder of the proof follows directly from the proof of Proposition 1 under the assumption that standard identifying assumptions for the ATT are satisfied. ■

Appendix B Formal Exposition of Truncation by Death Problem

Table A1 reproduces Table 1 using more standard parameterization of selection and outcomes from the truncation by death literature. Suppose that treatment $Z_i \in \{0, 1\}$, is assigned such that the probability of assignment to treatment $Z = 1$ is $p \in (0, 1)$ for all units. A first outcome, $S(Z) \in \{0, 1\}$ indicates whether the a subject "survives." The dependent variable of interest $Y(S, Z)$ occurs subsequent to the realization of $S(Z)$. Define four causal types (principal strata): always survivors, if treated survivors, if untreated survivors, and never survivors. Table A1 defines these types, their shares in the population, and relevant potential outcomes.

Stratum	Weight	$S(Z=1)$	$S(Z=0)$	$Y(S=1, Z=1)$	$Y(S=1, Z=0)$	$Y(S=0, Z=1)$	$Y(S=0, Z=0)$
Always survivor	π_A	1	1	$Y_A(1,1)$	$Y_A(1,0)$	-	-
If **T**reated survivor	π_T	1	0	$Y_T(1,1)$	-	-	$[Y_T(0,0)]$
If **U**ntreated survivor	π_U	0	1	-	$Y_U(1,0)$	$[Y_U(0,1)]$	-
Never survivor	π_N	0	0	-	-	$[Y_N(0,1)]$	$[Y_N(0,0)]$

Table A1: Principal strata of an experiment with a binary treatment and binary survival variable. Elements in brackets indicate that a potential outcome is undefined. If defined, the outcome $Y(S, Z) \in \mathbb{R}^1$ and the last four columns indicate cell means.

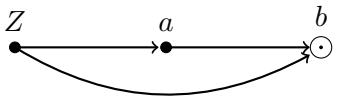
Given the binary assignment to treatment and the binary survival variable, the ATE of Z on Y could ideally be written:

$$\begin{aligned} E[Y(Z=1)] - E[Y(Z=0)] =& \pi_A E[Y_A(1,1)] + \pi_T E[Y_T(1,1)] + \pi_U \underline{E[Y_U(0,1)]} + \pi_N \underline{E[Y_N(0,1)]} - \\ & (\pi_A E[Y_A(1,0)] + \pi_T \underline{E[Y_T(0,0)]} + \pi_U E[Y_U(1,0)] + \pi_N \underline{E[Y_N(0,0)]}) \end{aligned} \tag{10}$$

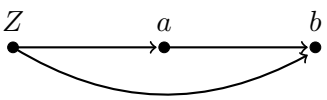
However, because some of these quantities (underlined in red) are undefined, the expression (and hence the ATE) is undefined.

Appendix C Directed Acyclic Graph Representation

Truncation by death, and analogous problems in social science, can also be depicted in directed acyclic graphs. Figure A1 depicts two possible relationships between a randomly assigned treatment, Z, and two sequential outcomes, a and b. Panels (a) and (b) correspond directly to the game trees in Figure 1. In both panels, b is a function of both Z and a, denoted $b(Z, a)$, whereas a is only a function of Z, denoted $a(Z)$. Unlike in the right panel, in the left panel, the node denoted by $\odot$ indicates that b is defined for only some values of a. In other words, there exist some units for which the potential outcome $b(Z, a)$ is not defined.



(a) Implied by strategy set asymmetry



(b) Implied by strategy set symmetry

Figure A1: Two graphical causal models of a randomly-assigned treatment, Z, and two sequential outcomes, a and b. The $\odot$ node indicates that outcome variable b is not defined for all levels of outcome variable a.

Appendix D Equilibrium Characterization, Proofs from Stylized Models

A4.1 Game Trees for Other Cases

Figure 3 depicts the game tree for Case #4 of the model. Figure A2 depicts the game trees for the three (nested) cases of the model listed in Table 2.

A4.2 Case #1: Always Crime

In this decision theoretic model, I assume that a crime occurred with probability 1. The bystander reports if the expected utility from reporting $E[U_B(r)]$ exceeds the expected utility from not reporting $E[U_B(\neg r)]$:

$$E[U_B(r)] \geq E[U_B(\neg r)] \Rightarrow \iota_R \psi - c_R \geq \iota_N \psi$$

Solving for ψ, the citizen will report if:

$$\psi \geq \frac{c_R}{\iota_R - \iota_N} \qquad \blacksquare$$

Given $F_\psi(\cdot)$, the cdf of ψ is drawn, the proportion of citizens that report a crime is $1 - F_\psi\left(\frac{c_R}{\iota_R - \iota_N}\right)$. With this rate of reporting, the ATE on reporting can be written:

$$\begin{aligned} ATE^1_{\mathcal{R}} &= 1 - F_\psi\left(\frac{c_R^{Z=1}}{\iota_R - \iota_N}\right) - \left(1 - F_\psi\left(\frac{c_R^{Z=0}}{\iota_R - \iota_N}\right)\right) \\ &= F_\psi\left(\frac{c_R^{Z=0}}{\iota_R - \iota_N}\right) - F_\psi\left(\frac{c_R^{Z=1}}{\iota_R - \iota_N}\right) > 0 \end{aligned}$$

This quantity is positive because $c_R^{Z=1} < c_R^{Z=0}$. Further, the ATE on incidence in the administrative record is:

$$\begin{aligned} ATE^1_{\mathcal{V}} &= \iota_R \underbrace{\left[1 - F_\psi\left(\frac{c_R^{Z=1}}{\iota_R - \iota_N}\right)\right]}_{\text{Reporting rate}} + \iota_N \underbrace{\left[F_\psi\left(\frac{c_R^{Z=1}}{\iota_R - \iota_N}\right)\right]}_{\text{Non-reporting rate}} - \iota_R \underbrace{\left[1 - F_\psi\left(\frac{c_R^{Z=0}}{\iota_R - \iota_N}\right)\right]}_{\text{Reporting rate}} - \iota_N \underbrace{\left[F_\psi\left(\frac{c_R^{Z=0}}{\iota_R - \iota_N}\right)\right]}_{\text{Non-reporting rate}} \\ &= (\iota_R - \iota_N)\left[F_\psi\left(\frac{c_R^{Z=0}}{\iota_R - \iota_N}\right) - F_\psi\left(\frac{c_R^{Z=1}}{\iota_R - \iota_N}\right)\right] > 0 \end{aligned}$$

This quantity is positive because $c_R^{Z=1} < c_R^{Z=0}$ and $\iota_R > \iota_N$. Because crime always occurs (there is no selection), the ATE is equivalent to the SACE in both cases. $\blacksquare$

A4.3 Case #2: Exogenous Crime and Exogenous Investigation

This model directly follows from Case #1. However, in the $1-\rho$ proportion of cases (precincts) in which there is no crime perpetrated, the reporting outcome is undefined. As such, $ATE_{\mathcal{R}}$ and $ATE_{\mathcal{V}}$ are undefined. In the ρ proportion of cases in which there is crime, the SACE follows from the calculation of the ATE from Model A4.2. Thus:

$$SACE_{\mathcal{R}} = F_\psi\left(\frac{c_R^{Z=0}}{\iota_R - \iota_N}\right) - F_\psi\left(\frac{c_R^{Z=1}}{\iota_R - \iota_N}\right) > 0$$
$$SACE_{\mathcal{V}} = (\iota_R - \iota_N)\left[F_\psi\left(\frac{c_R^{Z=0}}{\iota_R - \iota_N}\right) - F_\psi\left(\frac{c_R^{Z=1}}{\iota_R - \iota_N}\right)\right] > 0$$

Now consider the quantities estimated by a difference-in-means, $\Delta_{\mathcal{R}}$ and $\Delta_{\mathcal{N}}$:

$$\begin{aligned}\Delta_{\mathcal{R}} &= \rho\left(1 - F_\psi\left(\frac{c_R^{Z=1}}{\iota_R - \iota_N}\right)\right) - \rho\left(1 - F_\psi\left(\frac{c_R^{Z=0}}{\iota_R - \iota_N}\right)\right) \\ &= \rho SACE_{\mathcal{R}} \\ \Delta_{\mathcal{V}} &= (\iota_R - \iota_N)\left[\rho F_\psi\left(\frac{c_R^{Z=0}}{\iota_R - \iota_N}\right) - \rho F_\psi\left(\frac{c_R^{Z=1}}{\iota_R - \iota_N}\right)\right] \\ &= \rho SACE_{\mathcal{V}}\end{aligned}$$

A naive comparison of treatment and control beats will yield the quantities $\rho SACE_{\mathcal{R}}$ and $\rho SACE_{\mathcal{V}}$, respectively. Both quantities are positive. ■

A4.4 Case #3: Endogenous Crime and Exogenous Investigation

I characterize a subgame perfect equilibrium in pure strategies by backward induction. As such, I begin with the citizen's decision whether or not to report a crime in the subgame in which a crime has occurred. This is equivalent to the citizen's calculation in subsection A4.2. The citizen reports if and only if:

$$\psi \geq \frac{c_R}{\iota_R - \iota_N}$$

Now consider the suspect's choice. He will commit a crime if the expected utility from reporting $E[U_S(v)]$ exceeds the expected utility from not reporting $E[U_S(\neg v)]$:

$$\begin{aligned}\lambda - p\left[\iota_R\left[1 - F_\psi\left(\frac{c_R}{\iota_R - \iota_N}\right)\right] + \iota_N F_\psi\left(\frac{c_R}{\iota_R - \iota_N}\right)\right] &\geq 0 \\ p\left[\iota_R + (\iota_N - \iota_R)F_\psi\left(\frac{c_R}{\iota_R - \iota_N}\right)\right] &\leq \lambda\end{aligned}$$

In the unique subgame perfect equilibrium, thus, the suspect commits a crime if:

$$\lambda \geq p\left[\iota_R + (\iota_N - \iota_R)F_\psi\left(\frac{c_R}{\iota_R - \iota_N}\right)\right]$$

Upon observing the crime, the bystander reports if $\psi \geq \frac{c_R}{\iota_R - \iota_N}$. ■

As in the previous case, the ATEs are undefined because some crimes do not occur. To compute the SACE, first it is useful to define two thresholds of λ which define crime occurrence under treatment and control:

$$\tilde{\lambda} = p\left[\iota_R + (\iota_N - \iota_R)F_\psi\left(\frac{c_R^{Z=1}}{\iota_R - \iota_N}\right)\right]$$
$$\underset{\sim}{\lambda} = p\left[\iota_R + (\iota_N - \iota_R)F_\psi\left(\frac{c_R^{Z=0}}{\iota_R - \iota_N}\right)\right]$$

Because $c_R^{Z=0} > c_R^{Z=1}$, $\tilde{\lambda} > \underset{\sim}{\lambda}$. This implies that any crime that would occur if a unit is treated would occur if the unit is untreated. The "always survivor" stratum is thus defined by any suspect for whom $\lambda > \tilde{\lambda}$. The SACEs are thus given by:

$$SACE_{\mathcal{R}} = F_\psi\left(\frac{c_R^{Z=0}}{\iota_R - \iota_N} \mid \lambda > \tilde{\lambda}\right) - F_\psi\left(\frac{c_R^{Z=1}}{\iota_R - \iota_N} \mid \lambda > \tilde{\lambda}\right) > 0$$
$$SACE_{\mathcal{V}} = (\iota_R - \iota_N)\left[F_\psi\left(\frac{c_R^{Z=0}}{\iota_R - \iota_N} \mid \lambda > \tilde{\lambda}\right) - F_\psi\left(\frac{c_R^{Z=1}}{\iota_R - \iota_N} \mid \lambda > \tilde{\lambda}\right)\right] > 0$$

A difference-in-means estimator estimates:

$$\begin{aligned}
\Delta_{\mathcal{R}} =& F_\lambda(\tilde{\lambda})\left[F_\psi\left(\frac{c_R^{Z=0}}{\iota_R - \iota_N} \mid \lambda > \tilde{\lambda}\right) - F_\psi\left(\frac{c_R^{Z=1}}{\iota_R - \iota_N} \mid \lambda > \tilde{\lambda}\right)\right] - \\
& (F_\lambda(\tilde{\lambda}) - F_\lambda(\underset{\sim}{\lambda}))F_\psi\left(\frac{c_R^{Z=0}}{\iota_R - \iota_N} \mid \lambda \in (\underset{\sim}{\lambda}, \tilde{\lambda}]\right) \\
=& SACE_{\mathcal{R}} - (F_\lambda(\tilde{\lambda}) - F_\lambda(\underset{\sim}{\lambda}))F_\psi\left(\frac{c_R^{Z=0}}{\iota_R - \iota_N} \mid \lambda \in (\underset{\sim}{\lambda}, \tilde{\lambda}]\right) \\
\Delta_{\mathcal{V}} =& F_\lambda(\tilde{\lambda})\left[(\iota_R - \iota_N)\left[F_\psi\left(\frac{c_R^{Z=0}}{\iota_R - \iota_N} \mid \lambda > \tilde{\lambda}\right) - F_\psi\left(\frac{c_R^{Z=1}}{\iota_R - \iota_N} \mid \lambda > \tilde{\lambda}\right)\right]\right] - \\
& (F_\lambda(\tilde{\lambda}) - F_\lambda(\underset{\sim}{\lambda}))\left[(\iota_R - \iota_N)F_\psi\left(\frac{c_R^{Z=0}}{\iota_R - \iota_N} \mid \lambda \in (\underset{\sim}{\lambda}, \tilde{\lambda}]\right)\right] \\
=& SACE_{\mathcal{V}} - (F_\lambda(\tilde{\lambda}) - F_\lambda(\underset{\sim}{\lambda}))\left[(\iota_R - \iota_N)F_\psi\left(\frac{c_R^{Z=0}}{\iota_R - \iota_N} \mid \lambda \in (\underset{\sim}{\lambda}, \tilde{\lambda}]\right)\right]
\end{aligned}$$

The sign of the difference-in-means estimator is ambiguous for both outcomes. While both SACEs are positive, the second term in both expressions is negative. ■

A4.5 Case #4: Strategic Policing

I characterize a subgame perfect equilibrium in pure strategies by backward induction. As such, I begin with the officer's decision of whether or not to investigate a crime, conditional on whether the crime was reported:

$$\begin{aligned}
E[u_O(i|r)] &\geq E[u_O(\neg i|r)] \\
-\kappa &\geq -\alpha_r \\
\kappa &\leq \alpha_r
\end{aligned}
\qquad\qquad
\begin{aligned}
E[u_O(i|\neg r)] &\geq E[u_O(\neg i|\neg r)] \\
-\kappa &\geq -\alpha_{\neg r} \\
\kappa &\leq \alpha_{\neg r}
\end{aligned}$$

When the bystander evaluates the likelihood of reporting, the probability that a crime is investigated is thus given by $1 - F_\kappa(\alpha_r)$ (if reported) and $1 - F_\kappa(\alpha_{\neg r})$. Plugging these into the bystander's expected utility, the bystander reports if and only if:

$$\psi \geq \frac{c_R}{F_\kappa(\alpha_{\neg r}) - F_\kappa(\alpha_r)}$$

Now consider the suspect's choice. He will commit a crime if the expected utility from reporting $E[U_S(v)]$ exceeds the expected utility from not reporting $E[U_S(\neg v)]$. Denote the threshold above which a crime occurs as $\hat{\lambda}$.

$$\lambda - p\left[(1 - F_\kappa(\alpha_r))\left[1 - F_\psi\left(\frac{c_R}{F_\kappa(\alpha_{\neg r}) - F_\kappa(\alpha_r)}\right)\right] + (1 - F_\kappa(\alpha_{\neg r}))F_\psi\left(\frac{c_R}{F_\kappa(\alpha_{\neg r}) - F_\kappa(\alpha_r)}\right)\right] \geq 0$$

$$\hat{\lambda} \geq p\left[1 - F_\kappa(\alpha_r) + (F_\kappa(\alpha_r) - F_\kappa(\alpha_{\neg r}))\, F_\psi\left(\frac{c_R}{F_\kappa(\alpha_{\neg r}) - F_\kappa(\alpha_r)}\right)\right]$$

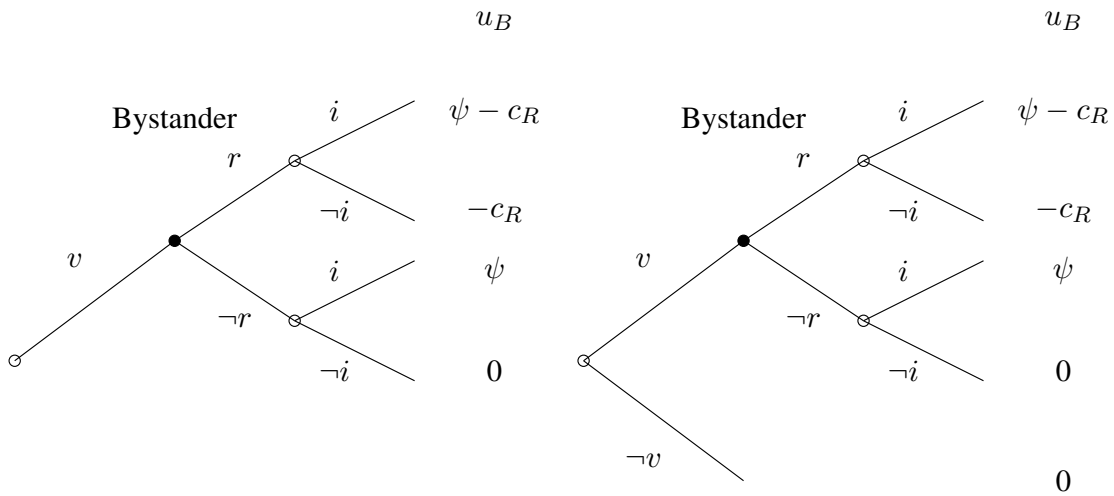
In the unique subgame perfect equilibrium, thus, the suspect commits a crime iff $\lambda > \hat{\lambda}$. Upon observing the crime, the bystander reports if $\psi \geq \frac{c_R}{\iota_R - \iota_N}$; and upon recieving the report, the officer investigates if $\kappa \leq \alpha_r$ but upon not receiving the report, the officer investigates iff $\kappa \leq \alpha_{\neg r}$. ■

This case is identical to the previous case except that $\iota_R \equiv 1 - F_\kappa(\alpha_R)$ and $\iota_N \equiv 1 - F_\kappa(\alpha_{\neg R})$. Substituting these expressions and redefining $\tilde{\lambda}$ and $\underset{\sim}{\lambda}$ as:

$$\tilde{\lambda} = p\left[1 - F_\kappa(\alpha_r) + (F_\kappa(\alpha_r) - F_\kappa(\alpha_{\neg r}))\, F_\psi\left(\frac{c_R^{Z=1}}{F_\kappa(\alpha_{\neg r}) - F_\kappa(\alpha_r)}\right)\right]$$

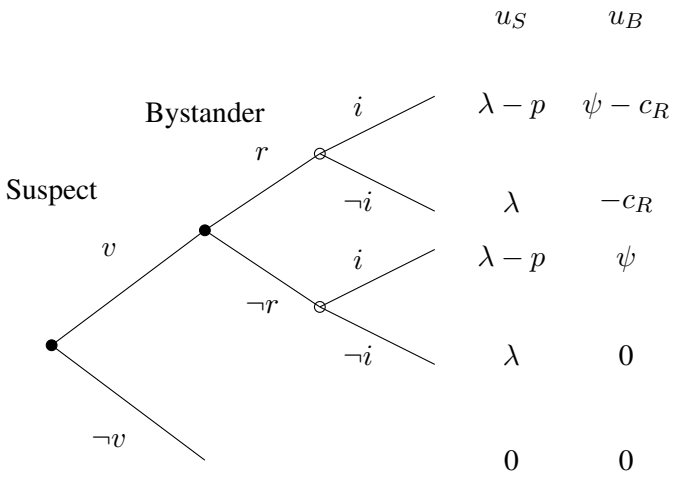
$$\underset{\sim}{\lambda} = p\left[1 - F_\kappa(\alpha_r) + (F_\kappa(\alpha_r) - F_\kappa(\alpha_{\neg r}))\, F_\psi\left(\frac{c_R^{Z=0}}{F_\kappa(\alpha_{\neg r}) - F_\kappa(\alpha_r)}\right)\right]$$

the remark follows directly from the proof to Remark 3. ■



CASE #1

CASE #2



CASE #3

Figure A2: Game trees corresponding to the other cases in Table 2.

Appendix E Applications in Social Science

This section expands upon the examples of truncation by death in the social sciences cited in Table 3. Some of these selection problems are discussed in the respective literature.

Table A2: Elaboration of truncation by death in literatures in Table 3

	Example	Truncation by death analogue	Mapping to framework
1	Effects of conflict	Individuals (militants or civilians) perish in conflict. This is quite literally truncation by death as in the medical setting.	The composition of actors changes between treatment (exposure to conflict) and outcome measurement. The potential outcomes measuring attitudes and behaviors are undefined for deceased subjects. Further, if conflict is deadly, the composition of a is arguably endogenous to the treatment.
2	Downstream effects of shocks on political behavior.	There are various possible shocks. In the cited example, Hall, Huff, and Kuriwaki (2019), recipients of a 19th century land lottery in Georgia had, on average, more children than non-winners of the land lottery. (Alternatively, recipients of the land lottery had more surviving children.)	The composition of actors changes between treatment (exposure to the wealth shock) and outcome measurement through differential rates of procreation. The potential outcomes measuring the behavior of children who would have been born (or survived) if their parents were treated are undefined.
3	Long-run effects of historical institutions.	Communities cease to exist or form after the imposition of (pre)-colonial institutions. While the prevalence of these issues are somewhat unclear, discussions of the inexact mapping between historic and current communities generally suggests these issues are present in multiple settings.	When communities are actors, the composition of actors changes between treatment (imposition of [pre]-colonial) institutions and outcome measurement. The potential outcomes measuring community-level behavioral outcomes in communities that ceased to survive are undefined.
4	Email audit experiments	Subjects choose whether or not to respond to an email.	Suppose the subject can provide an accurate or inaccurate response to the query. Subsequent to a decision to respond to the email, the response quality strategy set is: $S_a = \{\text{accurate}, \text{inaccurate}\}$. Subsequent to a decision not to respond, $S_a = \emptyset$. As such the potential outcomes measuring the quality of information provided is undefined if the subject chooses not to respond.
5	Ideological positioning across elections as a function of electoral performance at time t.	Between elections t and $t+1$, a party disbands. One could view this as a change in the set of parties (actors) at election $t+1$ or a change in the strategies available to parties that are contesting elections vs. disbanded.	A party that chooses to disband does not choose a platform in election $t+1$. As such, the potential outcomes measuring platform choice in $t+1$ (or functions thereof) are undefined.
6	Incumbency (dis)advantage	Between election t and election $t+1$ a candidate chooses not to seek re-election. As such, voters in $t+1$ do not have the choice to vote for a candidate who is not contesting office.	The voter's strategy set – measuring the options on the ballot– is different if the incumbent and challenger contest office in $t+1$ than if they do not. If the incumbent does not run, for example, the potential outcome measuring voter's decision to re-elect the incumbent is undefined.
7	Police use of force	A police officer chooses not to stop a civilian. The decision to use force (subsequent to a stop) depends on whether or not a civilian was stopped or not.	The strategies available to an officer, S_a, are different subsequent to stopping a civilian versus not stopping a civilian. As such the potential outcome measuring police use of force is undefined if the citizen is not stopped.

Table A3: Guide to implementing strategies for addressing phantom counterfactuals.

Strategy	Scope for application	Pros	Cons
1. Estimate only defined and identified estimands.	• Wide. Can be implemented in any study with phantom counterfactuals. • (However, note limits to estimation of the SACE.)	• Wide application. • Does not require any changes to treatments or the collection of additional data.	• SACEs are hard to estimate and existing estimators do not yet generalize to all research designs affected by phantom counterfactuals, see Table A4. • SACEs measure a "partial equilibrium" effect of treatment, which may or may not be informative about general equilibrium effects.
2. Re-randomize at histories where selection occurs.	• Very narrow. • Possible only in contemporaneous studies where (re-)randomization is feasible and ethical.	• Allows for straightforward estimation of a related ATE on outcomes after post-treatment selection.	• Ancillary experimentation is often very costly to implement and is impossible in many settings with phantom counterfactuals. • ATE of an ancillary treatment measures a "partial equilibrium" effect of a related treatment, which may not be informative about general equilibrium effects.
3. Change the set of outcomes.	• Intermediate. • Can be implemented in many experimental and observational studies. • For historical studies or re-analysis of existing studies, it may not be possible to measure necessary outcomes.	• Allows for identification of additional ATEs before post-treatment selection or (possibly) related ATEs after post-treatment selection.	• Collection of additional outcome data may be expensive or infeasible. • Redefined outcomes can be difficult to interpret. Treatment effects on these outcomes may not reflect theory- or policy-relevant quantities.

Table A4: Application of research design strategies in Table 4 to literatures in Table 3

	Example and cite	Estimate identified estimands	Re-randomize ancillary treatment	Redefine outcomes
1	Effects of conflict: (Blattman, 2009)	**Feasible:** The ATE of child soldiering on survival can be estimated from the survey data, though it is not reported.* The SACE of child soldiering on political participation can similarly be estimated, possibly using an interval estimator (Zhang and Rubin, 2003) or sensitivity analysis (Knox, Lowe, and Mummolo, 2020).	**Infeasible:** An ancillary randomization related to child soldiering is infeasible and unethical.	**Feasible**: For voting outcomes (currently a binary indicator) one flatten these measures into a categorical outcome consisting of $\{$Perished in conflict, Voted, Did not vote$\}$ to estimate ATEs on different categories. Note that interpretation of these ATEs is non-trivial.
2	Downstream effects of shocks on political behavior. (Hall, Huff, and Kuriwaki, 2019)	**Feasible:** It is possible to report the ATE of winning the land lottery on family size/composition. Given that the selection outcome is given by family size $\in \mathbb{N}$ (as opposed to binary), one would need to adapt current SACE estimators that are focused on binary outcomes to estimate a related estimand.†	**Infeasible:** Selection occurred over 150 years ago. It is infeasible to re-randomize a historical wealth.	**Infeasible:** It is hard to redefine or flatten outcomes given that the composition of units is endogenous to treatment. (In problems of truncation by death, we know the ex-ante set of units.)
3	Long-run effects of historical institutions. (Jha, 2013)	**Feasible**: It is possible to estimate the ATE of a treatment on persistence of a historical community given records of historical communities and current communities. Using bounding estimators or sensitivity analysis, one could estimate the SACE of colonial institutions (here medieval ports) on downstream inter-ethnic violence.	**Infeasible**: Medieval ports were designated around the 8th century AD. It is not possible to re-randomize historic port location at present.	**Potentially feasible**: Given a clear mapping between historical and present communities, and downstream community violence pre- and post-independence, one could define a community-level categorical outcome such as: $Y \in \{$No present community, Community with low inter-ethnic violence, Community with high inter-ethnic violence$\}$. Caution is necessary in interpreting ATEs on different categorical outcomes, but identification should be feasible given a mapping from historic to present communities.
4	Email audit experiments (White, Nathan, and Faller, 2015)	**Feasible**: Many authors estimate the ATE of petition content (petitioner identity) on rates of response. It is also possible to estimate a SACE on response quality with off-the-shelf estimators such as Zhang and Rubin (2003), among others.	**Feasible**: A researcher could randomly assign the content of a petitioner's follow-up email to an audit subject, conditional on response to the original email. Conditioning on response may increase the rate of response to the follow-up email, reducing the threat of phantom counterfactuals in the second (ancillary) experiment.	**Feasible**: Like Findley, Nielson, and Sharman (2014), researchers could flatten responses into a categorical response including (a) non-response; and (b) categorical classifications of response content.

* Blattman (2009) pools non-survival and out-migration as forms of attrition. He measures outcomes for migrants by interviewing surviving family members. Political participation outcomes are not undefined for migrants but they are undefined for those that are presumed dead in conflict. Kidnapped child soldiers are presumed dead in conflict at higher rates.

† The survivor stratum to which the SACE refers assumes that selection is binary. When selection is discrete but not necessarily binary (as in the number of children), one could specify a related estimand. For example, a stratum could be the families that would have 1 son regardless of the lottery outcome. Discrete-valued selection admits more strata and will place greater demands on the data.

	Example and cite	Estimate identified estimands	Re-randomize ancillary treatment	Redefine outcomes
5	Ideological positioning across elections as a function of electoral performance at time t. (Adams, 2012)	**Feasible**: One could estimate the ATE or ATT of whether a party survives to contest office in $t+1$.† One could estimate the SACE using existing estimators or develop estimators to estimate the survivor ATT.	**Infeasible**: An ancillary experiment would require the randomization of a new, related treatment between the time that parties choose to contest office at $t+1$ and their announcement of platforms. This is challenging for two reasons. First, designing a randomizable treatment that speaks to past electoral performance is quite challenging. Second, in order to sufficiently power statistical analyses, scholars studying ideological positioning often pool over countries and time (elections). This increases the practical challenges of conducting a secondary experiment.	**Potentially feasible**: One could measure other articulations of ideology – i.e., latent variable scalings of MPs' voting behavior or party members' communications – prior to the determination of candidacy, though note that these are not necessarily platforms. The feasibility of these approaches will depend somewhat on what can be inferred from behavior under different legislative institutions. Note however, that these measures of ideology are distinct from party platforms.
6	Incumbency (dis)advantage (Klašnja and Titunik, 2017)	**Feasible**: One could straightforwardly estimate the ATE of incumbency on contesting election in $t+1$. It is straightforward to estimate the SACE of incumbency on vote share in $t+1$ among candidates that "always run" using existing bounding or sensitivity analysis estimators.	**Infeasible**: It is not clear how one could feasibly or ethically re-randomize incumbency during the campaign period (after candidates declare candidacy).	**Feasible**: One could measure voter support for the incumbent *before* candidacy for election $t+1$ is announced. This is a different outcome, but it relates to voter support for incumbents and presents prior to candidate selection into electoral contests at $t+1$.
7	Police use of force (Knox, Lowe, and Mummolo, 2020)	**Infeasible**: Knox, Lowe, and Mummolo (2020) provide sensitivity analysis for the SACE of race on police use of force. They do not analyze the ATE of race on police stops.	**Infeasible**: It is not feasible to re-randomize race after a police has stopped a citizen (after observing their race).	**Unlikely to be feasible**: In the absence of our ability to observe all potential police-citizen encounters, redefining potential outcomes is quite challenging. If we were able to identify and measure the full set of potential police-citizen encounters (i.e., through continuous body camera footage), one could define a categorical potential outcome such as $Y \in \{$No contact, Stopped without use of force, and Stopped with use of force$\}$ and estimate ATEs on each category of "encounter."

† This literature is comparatively less attentive to concerns of causal identification than others in this table, in part because of the challenges posed by this data. The assumptions required to estimate an ATE or ATT on a party contesting election $t+1$ may be too strong for many practitioners of design-based causal inference.

Supplementary Appendix: References

Adams, James. 2012. "Causes and Electoral Consequences of Party Policy Shifts in Multiparty Elections: Theoretical Results and Empirical Evidence." *Annual Review of Political Science* 15: 401–419.

Angrist, Joshua D., and Jorn-Steffen Pischke. 2010. "The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics." *Journal of Economic Perspectives* 24 (2): 3–30.

Blattman, Christopher. 2009. "From Violence to Voting: War and Political Participation in Uganda." *American Political Science Review* 103 (2): 231–247.

Findley, Michael G., Daniel L. Nielson, and J.C. Sharman. 2014. "Causes of Noncompliance with International Law: A Field Experiment on Anonymous Incorporation." *American Journal of Political Science* 59 (1): 146–161.

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

Hall, Andrew B., Connor Huff, and Shiro Kuriwaki. 2019. "Wealth, Slaveownership, and Fighting for the Confederacy: An Empirical Study of the American Civil War." *American Political Science Review* 113 (3): 658–673.

Jha, Saumitra. 2013. "Trade, Institutions, and Ethnic Tolerance: Evidence from South Asia." *American Political Science Review* 107 (4): 806–832.

Klašnja, Marko, and Rocío Titunik. 2017. "The Incumbency Curse: Weak Parties, Term Limits, and Unfulfilled Accountability." *American Political Science Review* 111 (1): 129–148.

Knox, Dean, Will Lowe, and Jonathan Mummolo. 2020. "Administrative Records Mask Racially Biased Policing." *American Political Science Review* 114 (3): 619–637.

White, Ariel R., Noah L. Nathan, and Julie K. Faller. 2015. "What Do I Need to Vote? Bureaucratic Discretion and Discrimination by Local Election Officials." *American Political Science Review* 109 (1): 129–142.

Zhang, Junni L., and Donald B. Rubin. 2003. "Estimation of Causal Effects via Principal Stratification When Some Outcomes are Truncated by "Death"." *Journal of Educational and Behavioral Statistics* 28 (4): 353–368.