Phantom Counterfactuals

Tara Slough*

March 4, 2022

Abstract

Researchers often seek to identify the effects of a treatment on a sequence of behaviors, such as whether citizens register to vote and whether they then cast ballots. I show that average treatment effects (ATEs) are only identified until the first behavior (registering to vote) that affects the set of possible subsequent actions (voting). When one action changes the set of possible subsequent actions, it creates “phantom counterfactuals,” or undefined potential outcomes, which render ATEs unidentified. I show that applied theory allows researchers to diagnose phantom counterfactuals, which helps to recognize unidentified ATEs and focus instead on other estimands that are identified. I illustrate this approach using a stylized model of crime reporting, showing how different theories generate different sets of identified estimands while holding constant an experimental design. I thereby establish the necessity of applied theory for causal identification in empirical research with sequential behavioral outcomes.

Word count: 9839 words

*I thank Neal Beck, Alex Coppock, Scott de Marchi, Bob Erikson, Don Green, Justin Esarey, Daniel Hidalgo, Macartan Humphreys, Dorothy Kronick, Thomas Leavitt, Winston Lin, Kevin Munger, Fredrik Sävje, Mike Ting, Scott Tyson, Stephane Wolton, four anonymous reviewers, and audiences at the Harvard Experiments Working Group, Yale Quantitative Methods Seminar, Duke PPE Summer Symposium, Polmeth XXXVI, and the International Methods Colloquium for helpful comments. I am grateful to Scott Tyson for coining the name “phantom counterfactuals.” This project is supported in part by an NSF Graduate Research Fellowship, DGE-11-44155.

1 Introduction

Causal relationships are, by nature, sequential. Indeed, writers since at least Hume (1739-40 (2003)) and Kant (1781 (1996)) have noted that cause precedes effect. Sequence is so fundamental to causal inference that we often describe events as "pre-treatment" or "post-treatment," and the literature provides precise guidance about what timing-relative-to-treatment means for causal identification (e.g., Shadish, Cook, and Campbell, 2002). By comparison, another common sequencing issue – the sequence of actions taken in the post-treatment period, such as a citizen's decision to register to vote and her subsequent decision to cast a ballot – has received much less attention. In this paper, I show that the sequencing of post-treatment outcomes presents underappreciated limits to causal identification in social science. I also provide a framework for recognizing and addressing these limits.

Suppose that a researcher seeks to estimate the effect of a treatment on two sequential outcomes that measure behaviors (actions). If the realization of today's action changes the set of strategies available tomorrow – registering to vote grants the option of voting – then today's action renders tomorrow's potential outcomes undefined. I refer to these undefined potential outcomes as "phantom counterfactuals." Phantom counterfactuals are widespread across the social sciences. For example, voters cannot vote for an incumbent candidate when that candidate has chosen not to contest the next election (Erikson, 1971). Individuals' behaviors after conflict are undefined when they have died in conflict (Blattman, 2009). Police officers do not choose whether to use force against a citizen if they have not first stopped the citizen (Knox, Lowe, and Mummolo, 2020). Because the average treatment effect (ATE) is defined in terms of expectations evaluated over potential outcomes, phantom counterfactuals render the ATE undefined. An undefined estimand is not identified (Holland, 1986).

I show that applied theory allows for the diagnosis of phantom counterfactuals. By "applied theory," I mean simply writing down the extensive form of a theory of post-treatment behavior (i.e., a game tree). Phantom counterfactuals are present if and only if the branches of the tree are asymmetric, because asymmetry indicates a change in the subsequent actor or in the strategies available to that actor. Outcomes realized *after* a game tree first exhibits asymmetry are plagued by phantom counterfactuals (undefined potential outcomes), and the ATE on these outcomes is unidentified. An explicit theory of post-treatment behavior is therefore necessary for claims of causal identification when outcomes are sequential. Remarkably, a game tree alone—even in the absence of utilities or an equilibrium concept—is sufficient to evaluate these claims. The ATE on outcomes realized *before* the tree first exhibits asymmetry may be identified (under standard identification assumptions); those realized later are undefined and unidentified.

This argument generalizes the phenomenon of "truncation by death" in social science (e.g. Zhang and Rubin, 2003; McConnell, Stuart, and Devaney, 2008). The concept of truncation by death comes from clinical studies in which study participants could die between the time of treatment assignment and the time of outcome measurement. A study participant's death generates phantom counterfactuals for these outcomes of interest. Death itself represents one source of phantom counterfactuals relevant to social science, for example, in work on the effects of conflict. But phantom counterfactuals also affect many other, more mundane behavioral outcomes, spanning the voters' choices at the ballot box to email-based audit experiments.

To illustrate this link between theory and identification of the ATE on sequential outcomes, I develop a simple model of crime and policing. I consider an experiment that aims to reduce a bystander's cost of reporting crime, with the ultimate goal of increasing rates of reporting. Researchers aim to study crime reporting with administrative data: namely, 911 reports and measures of crime incidence. Through four variants of the model, I illustrate two implications of the general results. First, any selection into crime is analogous to death in clinical studies that exhibit truncation by death. It renders the ATE on subsequent outcomes—including 911 calls and administrative crime measures—undefined. We can see this simply by writing down the extensive form of a game in which the first move is a suspect's decision to commit or not commit a crime. After that initial node, the tree is asymmetric: only if the suspect commits a crime does the bystander have the opportunity to report it. This asymmetry alerts us to the presence of phantom counterfactuals and thus to the fact that the ATE on crime reporting and recording is undefined. Second, while other causal estimands – like the survivor average causal effect (SACE) – may be identified in the presence of phantom counterfactuals, standard estimators of the ATE do not, in general, recover the SACE. Moreover, when selection into crime is endogenous to the treatment – because suspects anticipate a higher likelihood of being caught when reporting increases – the resultant estimates cannot falsify any theoretical prediction.

As in this example, applied theory can identify problems of phantom counterfactuals but it also provides guidance on how researchers might address these problems empirically. I show that researchers can mitigate the identification problems associated with phantom counterfactuals by: (1) using applied theory to find a set of estimands that *are* identified, and focusing empirical effort on that set; (2) re-randomizing treatment; or (3) changing or redefining outcomes of interest. Critically, all of these recommendations follow from an explicit link between a minimal theoretical model and the empirical research design. For this reason, I advocate a closer marriage of theory and identification-oriented research designs.

This work makes three principal contributions to existing literature. First, I show that in a large class of research designs – those with sequential behavioral outcomes – theory is necessary for the identification of estimands like the ATE. This finding about the role of theory in establishing internal validity complements recent work on the necessity of theory for causal generalization (Gailmard, 2021) or external validity (Slough and Tyson, 2021). It also advances ongoing debates on the role of theory in identification-driven research (Clark and Golder, 2015; Ashworth, Berry, and Bueno de Mesquita, 2015; Samii, 2016; Huber, 2017; Franzese, 2020). Further, these results focused on identification complement previous discussions of the interpretation (or interpretability) of reduced-form causal estimands (Signorino, 2003; Signorino and Yilmaz, 2003; Heckman, 2008; Keane, 2010; Rust, 2010).

Second, my findings expand a growing literature on the "theoretical implications of empirical models" (TIEM) (e.g., Ashworth and de Mesquita, 2014; Eggers, 2017; Gailmard and Patty, 2018; Prato and Wolton, 2019; Izzo, Dewan, and Wolton, 2020). In particular, the identification problems posed by phantom counterfactuals constitute a new class of commensurability problems in which analysts aim to estimate a quantity that is theoretically undefined (Bueno de Mesquita and Tyson, 2020). I show how theory can locate identification problems and suggest remedies in research designs with sequential behavioral outcomes.

Finally, by focusing on the sequence of outcomes after treatment, I contribute to a growing literature on post-treatment selection in causal studies. Existing work largely focuses on bias induced by the inclusion of "bad" controls (Montgomery, Nyhan, and Torres, 2018); post-treatment sample conditioning (Aronow, Baron, and Pinson, 2019); or post-treatment selection in specific empirical applications (Knox, Lowe, and Mummolo, 2020; Coppock, 2019). This article unifies a general class of applications by focusing on identification instead of estimation. By linking phantom counterfactuals to dynamic models in political science, I expose the prevalence of this class of identification problems and provide a framework for addressing them.

2 From Extensive Form to Causal Identification

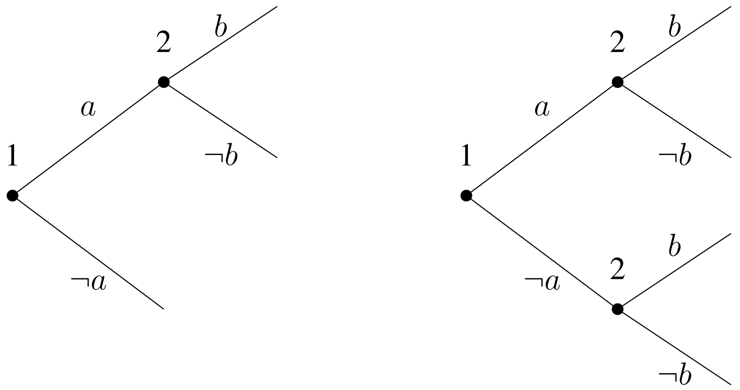
I focus on the identification of causal effects on *post-treatment behavioral outcomes*. Post-treatment behavioral outcomes measure actions that occur after the assignment of treatment. These outcomes often occur in sequence, rather than simultaneously. Sequential outcomes correspond to dynamic models.

Dynamic models specify the temporal sequence of possible actions. I use the word "history" to refer to the set of all previous post-treatment actions. Sequential outcome variables therefore measure actions or beliefs at different histories of a model. When describing dynamic models, I refer to the set of histories (nodes) as H, as is standard. The first post-treatment node is $H^{\emptyset}$. H^T represents a terminal node, or the last modeled post-treatment outcome. I define *strategy set symmetry*, which is useful for classifying post-treatment histories.

Definition 1. *Strategy set symmetry. A model exhibits strategy set symmetry if for any history, h, the subsequent actor, i, is the same and has an equivalent strategy set, S_i, regardless of the strategy selected at h, for all $h \in H \backslash H^T$.*

Strategy set symmetry is simple to visualize in a game tree. Figure 1 depicts two games. On the

Figure 1: Strategy set symmetry



ASYMMETRIC STRATEGY SETS

SYMMETRIC STRATEGY SETS

The game tree right exhibits strategy set symmetry. The game tree on the left does not.

left, Player 2's set of strategies, S_2, depends on the Player 1's action at the first node. If Player 1 plays a, Player 2's set of strategies is $S_2 = \{b, \neg b\}$. But if Player 1 plays $\neg a$, Player 2 does not play ($S_2 = \emptyset$). Player 2's strategy sets are asymmetric per Definition 1 because they depend on Player 1's action. In contrast, in the game in the right panel, Player 2's set of strategies, $S_2 = \{b, \neg b\}$ is equivalent regardless of Player 1's action. This means that the game on the right is strategy set-symmetric for both modeled actions.

While Figure 1 depicts a game theoretic model with two distinct actors, strategy set symmetry applies to many forms of dynamic models. Importantly, the actor at each history can be the same individual. This means that dynamic decision theoretic models, in which a single actor makes a sequence of decisions, can also be classified in this way.

Consider the mapping between the models in Figure 1 and a simple experiment. Suppose that an experiment seeks to compare the difference in the frequency with which a population of Player 1's chooses action a under some randomly-assigned $Z \in \{0, 1\}$, where 1 indicates assignment to treatment and 0 indicates assignment to control. In both panels, $a(Z)$ is defined for all units. As such, $E[a(Z = 1)] - E[a(Z = 0)]$ is also defined. Under standard identifying assumptions, this quantity is the ATE of the treatment Z on an outcome measuring action a.[1]

[1] In an experiment, these standard identifying assumptions are ignorability of treatment assign-

Now, suppose the researcher wants to understand the difference in the frequency with which a population of Player 2's chooses strategy b as a function of the same treatment. Denote potential outcomes $b(Z)$, where $b(Z) = 1$ corresponds to action b and $b(Z) = 0$ corresponds to action $\neg b$. In the left panel, this presents a problem. If player 1 plays a, then Player 2 plays b or $\neg b$, and the potential outcomes $b(Z)$ are defined. But if Player 1 plays $\neg a$, Player 2 does not act and their potential outcomes, $b(Z)$ are undefined. Expectations evaluated over these potential outcomes, $E[b(Z = 1)]$ and $E[b(Z = 0)]$ are similarly undefined and therefore the ATE of Z on b, $E[b(Z = 1)] - E[b(Z = 0)]$ is also undefined. One requirement for identification of causal estimands, including the ATE, is that all variables – including all potential outcomes – are defined for every unit in the experimental population (Holland, 1986). An undefined estimand is not identified.

On the other hand, in the right panel of Figure 1, Player 2 can play b or $\neg b$ regardless of Player 1's action. This means that the potential outcomes $b(Z)$ are defined at both histories, $H^1 \in \{a, \neg a\}$. In this case, ATE of Z on b is identified under standard identifying assumptions.

The difference in the identification of the ATE on Player 2's action across the two panels of Figure 1 reveals two critical insights. First, the same experiment can identify the ATE on some post-treatment outcome(s) but not others. In the left panel, the ATE of Z on a is identified but the ATE of Z on b is not. Second, the ATE of Z on b is identified on the basis of assumptions about the structure of the model. It is ultimately our theoretical assumption about whether we are in the left or the right panel that determines whether the ATE of Z on b is identified.

These findings based on the simple models in Figure 1 generalize to far more complex models of post-treatment behavior. In more complex dynamic models, identification of the ATE on a given outcome similarly depends on the strategy set symmetry of the model. If the model is not strategy set symmetric, only the ATE(s) of treatment on outcome(s) prior to the first strategy set asymmetric history are identified. Proposition 1 provides a general statement of this finding.[2]

ment, excludability, and the stable unit treatment value assumption (SUTVA).

[2] The proof of Proposition 1 considers a setting in which selection is represented as a binary choice or realization. The proof is also consistent with the common setting in which an actor's strategy set is continuous and her action is then mapped into a binary realization.

Proposition 1. *In an experiment in which standard identifying assumptions hold, 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 ATE is identified.*
2. *There exists at least one post-treatment behavioral outcome for which the ATE is not identified.*

In an experiment in which standard identifying assumptions hold, if a dynamic theory of post-treatment behavior is strategy set symmetric, then the ATE is identified for all modeled post-treatment behavioral outcomes. (All proofs in Appendix.)

Proposition 1 provides several insights. First, as in the discussion of Figure 1, the ATE is defined with respect to a specific *outcome*, not simply as a property of the empirical research design for any post-treatment variable. Recent emphasis identifying causal effects has often led to heavy focus on creating or "finding" exogenous variation via an experiment or natural experiment. The central challenge of the research design is therefore to find this variation; once located, these efforts can be leveraged to estimate the effects on a host of different post-treatment outcomes. The result identified here suggests that estimating the effects of the same treatment on multiple behavioral outcomes is not necessarily consistent with the stated motive of causal identification.

Second, the primary threat to identification of the ATE identified by Proposition 1 is post-treatment selection. Post-treatment selection occurs when the choice of a strategy at a history changes the menu of strategies available to the next actor. The first instance of post-treatment selection occurs at the first strategy set-asymmetric history of a model. The order of outcomes in a sequence is therefore crucial. The ATEs of treatment on outcomes before and inclusive of the first instance of post-treatment selection in a sequence are identified. After selection, the ATE is no longer identified.

Many works estimate ATEs on multiple outcomes but do not articulate an explicit theoretical model of the relationship between outcomes. Proposition 1 shows that when outcomes are sequential, the extensive form of an unstated model must support these claims to identification. Corollary

1 reveals what readers can learn about the unstated model from claims to identification of the ATE. These results suggest that the absence of theory should not be conflated with "agnosticism" in the presence of claims of identification.

Corollary 1. *In an experiment for which researchers claim to identify the ATE of more than one behavioral outcomes, it must be the case that the implied theoretical model*

(a) is not dynamic, or

(b) is dynamic and strategy-set symmetric for these outcomes.

To this point, the results emphasize limits to identification of the ATE in the presence of post-treatment selection. But other, less common, causal estimands are identified in the presence of post-treatment selection. With reference to the left panel of Figure 1, recall that the ATE is not identified because $b(Z)$ is undefined when Player 1 plays $\neg a$. By focusing only on the history $H^1 = \{a\}$, note that $b(Z)$ is defined for all units. However, we cannot directly condition on the realization of $a(Z)$, since this is a post-treatment outcome. Instead, we may seek to compare $E[b(Z)]$ for units for which $a(Z = 1) = 1$ and $a(Z = 0) = 1$. Substantively, this corresponds to interactions in which the first player first player would play a, regardless of whether they were treated or untreated. Note, however, that due to the fundamental problem of causal inference, we never observe both $a(Z = 1)$ and $a(Z = 0)$ for a unit. As such, membership in this subgroup is ultimately unobservable. The average causal effect among this subgroup, often called the survivor average causal effect (SACE) is defined:

$$\begin{aligned} SACE = E[b(Z = 1)|a(Z = 1) = 1, a(Z = 0) = 1]- \\ E[b(Z = 0)|a(Z = 1) = 1, a(Z = 0) = 1]. \end{aligned} \tag{1}$$

Proposition 2. *In an experiment in which standard identifying assumptions hold, if a dynamic theory of post-treatment behavior is not strategy set symmetric, then the SACE is identified for outcomes subsequent to the first strategy-set asymmetric history.*

Proposition 2 shows that, in the presence of post-treatment selection, the SACE is defined when the ATE is not. The proof builds on the intuition that $b(Z)$ is always defined when $a(Z) = 1$. This means that the SACE is identified on outcomes for which the ATE is not identified. While identification of the SACE is straightforward, estimation poses well-known challenges. These challenges are clear from examination of (1). Both expectations are conditioned on two potential outcomes, $a(Z = 1)$ and $a(Z = 0)$. However, as mentioned above, we never observe both potential outcomes for any given unit. Consistent with this intuition, in Propositions A1-A2 (p. A-4), I show that the difference-in-means estimator, which is often invoked to estimate the ATE, does not, in general, produce estimates that are informative about the sign or magnitude of the SACE.

2.1 Scope of results

It is useful to consider the scope of these results first in the context of a single causal process. To do so, I consider again the left panel of Figure 1, in which Player 1's decision of a or $\neg a$ determines the strategies available to Player 2. It is clear from Proposition 1 that ATE of Z on $a(Z)$ is identified. This identification does not imply substantive importance. For example, consider a treatment that encourages citizens to initiate a bureaucratic process like registering for a state ID or applying for a social program. Lodging the initial request for service may measure only compliance with treatment assignment. However, potential outcomes measuring subsequent interactions with state service providers are undefined when subjects do not request the service in the first place. Compliance with treatment assignment may or may not be an important outcome. As such, Proposition 1 does not guarantee that identified ATEs are substantively important.

To this point, I have focused on sequential behavioral outcomes. Yet researchers often measure treatment effects on beliefs or attitudinal outcomes in addition to behavioral outcomes. These attitudinal outcomes may evolve over time after treatment. Phantom counterfactuals present when the menu of possible beliefs or attitudes may be endogenous to post-treatment actions. I argue that these concerns are less prevalent in the study of attitudinal outcomes. In general, the set of attitudes or beliefs that can be expressed on a survey instrument do not vary by history. However, if post-treatment selection changes the composition of subjects that could feasibly express their attitudes or beliefs (i.e., through death or, in the longer term, through differential birth rates), identification challenges re-emerge. While these scenarios do present in some empirical settings, compositional changes – the source of post-treatment selection – are ultimately behavioral, not attitudinal.

2.2 How much theory?

Both panels of Figure 1 invoke a minimalist notion of a dynamic theory. These theoretical models include only a sequence of actors and their strategies. The results in Propositions 1 and 2 rely only on this minimal theoretical structure. As such, these identification results do not rely upon assumptions about actors' utilities, the information structure, the equilibrium concept, or any equilibrium refinements. Holding fixed the sequence of actors and strategies, the imposition of these additional theoretical assumptions cannot remedy – but similarly does not create – problems of ATE identification through phantom counterfactuals.

When we specify any theory, we make assumptions. Some of these assumptions are consequential for problems of identification while others are not. Corollary 1 shows that by making claims to identification of ATEs on multiple behavioral outcomes, researchers assume a minimal theoretical structure, whether stated or unstated. It is important to note that these assumptions related to sequence and strategy sets are comparatively observable. As a result, in some contexts, it may be possible to provide empirical support for these theoretical identifying assumptions.

My focus on identification is distinct from existing discussions focused on theory and the *interpretation* of causal estimands (Rust, 2010; Keane, 2010). When applied to causal estimands like the ATE or SACE, questions of interpretation focus on what these estimands measure in the underlying causal process. The theoretical assumptions that underpin identification of the ATE represent only a subset of those assumptions generally required to generate falsifiable predictions. For example, neither game tree in Figure 1 includes utilities or an equilibrium concept. As a result, neither figure generates a prediction about the sign of the ATE of treatment Z on a. Similarly, these trees do not generate predictions about the SACE of Z on b (left panel) or the ATE of Z on b (right panel). Additional theoretical assumptions are therefore critical to generating predictions and understanding whether causal estimands like ATEs and SACEs could falsify these predictions.

Do the strong assumptions needed for interpretation make claims to causal identification less credible? I argue that they generally do not. In settings with sequential behavioral outcomes – the focus of this paper – identification relies on only a subset of the theoretical assumptions needed for interpretation. Consider the assumptions imposed only for interpretation (e.g., players' utilities). So long as these assumptions do not contradict other "empirical" identification assumptions (i.e., SUTVA), they do not affect claims to identification. This has two important implications. First, misspecification of a theory is less of a threat to causal identification than interpretation. This occurs because identification relies on fewer, weaker, and often more testable assumptions. Second, reliance on additional theoretical assumptions to guide interpretation does not generally come at the expense of the credibility of identification.

3 The Link to Truncation by Death

The identification problem articulated above is analogous to the problem of truncation by death (e.g., Zhang and Rubin, 2003; McConnell, Stuart, and Devaney, 2008). In medical and epidemiological studies, truncation by death occurs when a subject dies after treatment but prior to the measurement of the ultimate outcome of interest. In these studies, researchers frequently seek to ascertain the quality of life under a new experimental therapy. However, if the patient dies before their quality of life measure is assessed, their relevant potential outcome(s) for the quality of life measure is undefined.

Table 1 illustrates the problem of truncation by death using principal stratification, following Frangakis and Rubin (2002) and Zhang and Rubin (2003). Maintaining the notation from Figure 1, first outcome "survival," denoted $a(Z)$, measures whether a patient survives until outcomes are measured, where $a(Z) = 1$ indicates survival and $a(Z) = 0$ indicates death. The strata are defined by both potential outcomes $a(Z = 1)$ and $a(Z = 0)$ for each unit. Of course, we can only observe one of these potential outcomes. Quality of life, denoted $b(Z, a)$ is measured *after* survival, and hence the outcomes are sequential. Importantly, these potential outcomes are a function of *both* treatment (Z) and survival ($a(Z)$). When a patient dies prior to the measurement of quality of life, their potential outcomes, $b(Z, a = 0)$ are undefined.

				$b(Z=1,$	$b(Z=0,$	$b(Z=1,$	$b(Z=0,$
Stratum	Share	$a(Z=1)$	$a(Z=0)$	$a=1)$	$a=1)$	$a=0)$	$a=0)$
Always survivor	π_A	1	1	$b_A(1,1)$	$b_A(0,1)$	-	-
If **T**reated survivor	π_T	1	0	$b_T(1,1)$	-	-	undefined
If **U**ntreated survivor	π_U	0	1	-	$b_U(0,1)$	undefined	-
Never survivor	π_N	0	0	-	-	undefined	undefined

Table 1: Principal strata assuming a binary treatment, Z. A first behavioral outcome akin to survival, $a(Z) \in \{0,1\}$, determines whether a second behavioral outcome $b(Z,a)$ is defined.

From Table 1, it is clear that the ATE of treatment on survival is identified under standard identifying assumptions, since $E[a(Z=1)] - E[a(Z=0)] = \pi_T - \pi_U$. In contrast, the ATE on quality of life, $E[b(Z=1)] - E[b(Z=0)]$, is undefined if any patient dies (if $\pi_A < 1$). This occurs because $E[b(Z=1)]$ and/or $E[b(Z=0)]$ is evaluated over at least one undefined potential outcome, and is therefore undefined. The SACE estimand described in Proposition 2 refers to the average causal effect of treatment on $b(Z,a)$ among the stratum of always survivors.

Undefined potential outcomes are distinct from attrition, or missing outcome data. Undefined potential outcomes are measured on a qualitatively different scale from defined potential outcomes (McConnell, Stuart, and Devaney, 2008). The difference in scales differentiates undefined outcomes from attrition or missingness. Indeed, in clinical studies, researchers are generally not missing outcome data – they observe that a subject died – but the observed outcome (death) is observed on a different scale from the outcome of interest (quality of life). Zhang and Rubin (2003: p. 353) formalize undefined potential outcomes that in problems of truncation by death, the ultimate outcome is valued on an extended space so $b(Z,a) \in \{\mathbb{R}, *\}$. Potential outcomes are undefined when $b(Z,a) = *$. If these outcomes were simply missing – rather than undefined – they would still be real-valued, even if they were unobserved.

The distinction between undefined outcomes and attrition is even clearer when considering statistical methods for missing data. First, consider multiple imputation (Rubin, 1987; King et al., 2001). In the context of truncation by death, multiple imputation could be used to impute quality of life measures for subjects that die. Yet, this implies a *loss* of information. We know that the subject died; imputing quality of life if the subject had lived provides a measure that is verifiably distinct from what occurred. Alternatively, consider resampling missing outcomes as a non-parametric alternative to imputation (Green and Gerber, 2012; Coppock et al., 2017). It is impossible to resample quality of life measures of deceased patients at least without changing some antecedent state of the world (keeping the patient alive). The mismatch between approaches for missing data and the inferential problems induced by truncation by death draw clear distinctions between the two pathologies in the context of research design.

The primary contribution of this article is to generalize the phenomenon of truncation by death by linking it to a common class of theoretical models in the social sciences. In the setting of truncation by death, both death and reporting of quality of life are behavioral outcomes. The first outcome – death – changes the strategy sets – or response options – available at the second outcome. This basic theoretical structure links truncation by death to many substantive problems in the social sciences.

In the remainder of this paper, I first illustrate these findings using stylized example of crime reporting and recording. Second, I describe instances of these identification problems across multiple literatures in political science. Finally, I provide suggestions for addressing or alleviating these concerns in empirical research.

4 Stylized Example

I now illustrate how different theories generate different sets of identified estimands within the same experiment. I write the theories formally so that I can compare estimates and causal estimands under each model. These theories are neither complex nor counterintuitive. Yet, the mapping between theoretical predictions and estimates is non-trivial even in these simple cases.

4.1 "See Something Say Something" and Crime Reporting: An Experiment

Consider a "see something, say something" campaign intended to increase crime reporting by citizens and, ideally, deter crime. The campaign is randomly assigned to police precincts within a city. Denote a binary treatment indicator, $Z_i \in \{0, 1\}$. Researchers measure outcomes using counts of geo-coded crime reports (e.g., 911 calls) aggregated to the precinct level, denoted $\mathcal{R}_i$, and geo-coded reported crime incidence data measured at the same level, denoted $\mathcal{V}_i$. Researchers seek to estimate the causal effect of the "see something, say something" messages on both outcomes. Suppose further that treatment assignment is ignorable, the treatment is excludable, and SUTVA holds.[3] In standard practice, researchers would generally seek to estimate the ATE of treatment on crime reporting and incidence. I consider the difference-in-means, Δ, which is estimated by (2).

$$\Delta = \overline{Y}(Z_i = 1) - \overline{Y}(Z_i = 0). \tag{2}$$

I compute difference-in-means estimates $\Delta_{\mathcal{R}}$ and $\Delta_{\mathcal{V}}$ for measured outcome variables $Y_i \in \{\mathcal{R}_i, \mathcal{V}_i\}$ under each model. I compare these quantities to the ATEs or SACEs derived from the same models. This comparison is important because the difference-in-means estimator is touted as an unbiased estimator of the ATE under the assumptions and experimental design that I have articulated (Angrist and Pischke, 2010; Green and Gerber, 2012).

4.2 Four Cases of a Model

I enumerate four cases of a simple model. These cases provide four accounts of the causal process underlying the reporting and crime recording outcomes of interest. Three features of these cases allow for direct comparability. First, I assume complete information in all cases. Second, I assume a common sequence of actions. Third, I use the same utilities. Collectively, these assumptions ensure comparability across both game theoretic and decision theoretic models. Among the game theoretic models, these assumptions allow me to use a common equilibrium concept.

I distinguish the goals of identification and interpretation when analyzing these cases. It is important to note that the identification findings rely on fewer assumptions than I elaborate when specifying the model. As above, they follow from examination of the game trees, even without utilities. However, interpretation of the resultant causal effects does rely on the full models, and thus invokes a set of assumptions about why players behave in the way that they do.

[3] The clustered treatment assignment at the neighborhood level in the present design is consistent with SUTVA under all models specified here.

The four cases of the model each assume some subset of three players: a bystander (B), a suspect (S), and an officer (O). S decides whether to commit a crime, denoted v or $\neg v$. By committing a crime, the suspect receives some surplus, $\lambda \geq 0$, drawn from the density $f_\lambda(\cdot)$ with cdf $F_\lambda(\cdot)$. However, if the suspect that commits the crime is investigated, she pays a penalty $p > 0$.

B observes whether a crime occurs. If the crime occurs, they chooses whether to report, at net cost $c_r > 0$. The "see something say something" treatment corresponds to a reduction in net costs of reporting, such that $c_r^{Z=1} < c_r^{Z=0}$, by providing information or appealing to social norms to report. If a crime is investigated, the bystander obtains a benefit, $\psi \geq 0$, conceived as a taste for order or justice. These tastes vary across the population and are drawn from the density $f_\psi(\cdot)$ with cdf $F_\psi(\cdot)$. I make no assumptions about the joint distribution of λ and ψ.

O observes that a crime occurred and whether it was recorded. They choose to investigate or not to investigate. An investigation requires some effort by the officer at cost $\kappa > 0$. κ is drawn from pdf $f_\kappa(\cdot)$. If they fail to investigate, they face an expected sanction of α. I assume that the officer is more likely to be sanctioned for failing to investigate reported crimes due to increased legibility of the crime such that: $0 < \alpha_{\neg r} < \alpha_r$.

The four cases of this model vary because different players are strategic. In all cases, the bystander decides whether to report a crime. Where any player is non-strategic, I parameterize the probability with which "nature" selects each strategy. Table 2 documents the relationship between the four models. The extensive form of the full model (Case #4) appears in Figure 2 and the extensive forms of the other cases appear in Figure A2 (p. A-11). As is clear in Figure 2, no reporting and no investigation occur if a crime has not occurred. As is clear, therefore, the occurrence of crime is the source of post-treatment selection in Cases #2-4.

Given complete information and the sequence of actions, I characterize the unique subgame perfect Nash equilibrium (SPNE) for both Cases #3 and #4. In the decision theoretic models (#1 and #2), I characterize the optimal behavior of the bystander. It is straightforward to solve these models so the analysis is relegated to Appendix D (p. A-7).

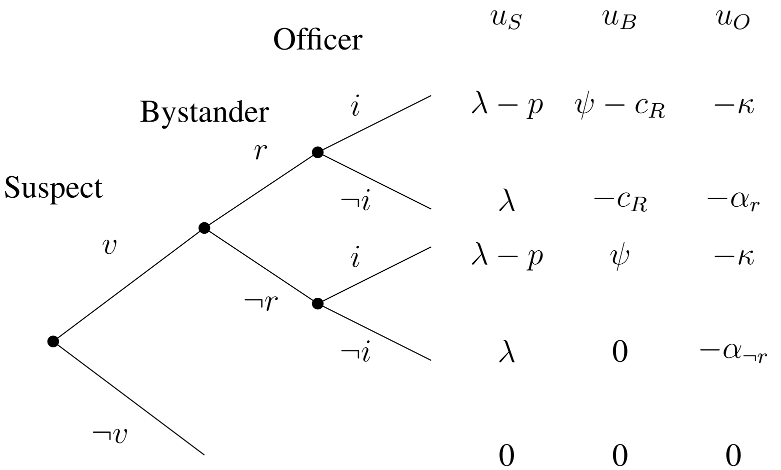
Mapping the model onto relevant causal estimands requires two additional considerations.

Case #1	**Case #2**
(1) A crime occurs with probability 1.	*(1) With probability, ρ, a crime occurs ("nature" commits a crime).*
(2) The bystander decides whether to report the crime.	(2) The bystander observes whether a crime was committed. If it was committed, she decides whether to report the crime.
(3) If a report is received, nature investigates with probability ι_R. If a report is not received, nature investigates with probability ι_N.	(3) If a report is received, nature investigates with probability ι_R. If a report is not received, nature investigates with probability ι_N.
(4) Utilities are realized.	(4) Utilities are realized.

Case #3	**Case #4**
(1) The suspect commits a crime or does not commit a crime.	(1) The suspect commits a crime or does not commit a crime.
(2) The bystander observes whether a crime was committed. If it was committed, she decides whether to report the crime.	(2) The bystander observes whether a crime was committed. If it was committed, she decides whether to report the crime.
(3) If a report is received, nature investigates with probability ι_R. If a report is not received, nature investigates with probability ι_N.	*(3) The officer observes whether a report was made and decides whether to investigate or not.*
(4) Utilities are realized.	(4) Utilities are realized.

Table 2: The sequence of the four cases of the model. The feature of each case emphasized in the discussion is italicized.

Figure 2: Model in case #4.



Extensive form representation of Case #4.

First, I define the mapping between actions in the model and the outcomes observed empirically. I assume that a bystander's reporting maps to the call data on reporting: $\mathcal{R}_i = 1$ if a crime occurs and it is reported. A case enters police records, $\mathcal{V}_i = 1$, if it is investigated by police (regardless of whether it was reported). Second, estimands are expressed in terms of expectations evaluated over the potential outcomes of multiple units. While the equilibria characterized correspond to an equilibrium occurrence of reporting or investigation in one precinct, I examine differences in average outcomes between treatment and control.

Case #1: Always Crime

In the simplest variant of the model, there is always a crime that the bystander could report, which implies that the game tree is strategy-set symmetric (see Figure A2, p. A-11). Here, we are only concerned with the bystander's decision of whether to report or not. The bystander will report if the costs of reporting are smaller than the expected benefit of restoring order. The ATE on reporting, then, is simply the difference in proportion of bystanders reporting the crime in treatment versus control beats. This quantity is positive since the net costs of reporting are lower in treatment than in control. Because rates of crime do not change in this model, when reporting goes up, police record more crimes, so the ATE on recording must also be positive. This interpretation follows from our assumptions about why the bystanders report crime. In the absence of selection into crime, the SACE and ATE are equivalent. Under the assumptions invoked here, the difference-in-means estimators are unbiased estimators of each ATE.

Remark 1. *When crime occurs with probability 1 (no selection), then:*

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

2. $ATE_{\mathcal{R}} = SACE_{\mathcal{R}}$ *and* $ATE_{\mathcal{V}} = SACE_{\mathcal{V}}$ *because there is no selection into crime.*

The quantities estimated by difference-in-means estimator on each outcome are: $\Delta_{\mathcal{R}} = ATE_{\mathcal{R}}$ *and* $\Delta_{\mathcal{V}} = ATE_{\mathcal{V}}$.

Case #2: Exogenous Crime

Case #2 introduces exogenous selection into crime. With probability $\rho \in (0,1)$ a crime occurs, independent of treatment of assignment. Because there are precincts with no crime, the bystander no longer faces the decision of whether to report in those precincts. As a result, ATE$_{\mathcal{R}}$ and ATE$_{\mathcal{V}}$ are no longer defined or identified. Importantly, this follows directly from the asymmetry in the game tree without any further assessment of the bystander's utility or decision. The relevant SACE estimands reflect the difference in rates of reporting and recording among precincts in which a crime would occur regardless of treatment assignment. In order to interpret why treatment should affect this SACE, we rely on assumptions about the bystander's utilities.

Even with *exogenous* selection, a naive difference-in-means no longer estimates the SACE. Since we do not observe true crime levels, this estimator effectively imputes an outcome of no reporting when crime does not occur. This equates non-reporting of crime that occurred with no crime occurrence. Since crime is exogenous, however, this estimator estimates the SACE scaled by the crime rate, ρ. With the present research design and data, ρ is not identified. However, the difference-in-means will maintain the same sign as the SACE. This is important if the goal is to evaluate the *sign* of the resultant treatment effect as a test of the theory.

Remark 2. *When crime occurs exogenously with probability $\rho \in (0,1)$, then:*

1. *$ATE_{\mathcal{R}}$ and $ATE_{\mathcal{V}}$ are undefined.*
2. $SACE_{\mathcal{R}} = F_\psi\left(\frac{c_R^{Z=0}}{\iota_R-\iota_N}\right) - F_\psi\left(\frac{c_R^{Z=1}}{\iota_R-\iota_N}\right) > 0$
   $SACE_{\mathcal{V}} = (\iota_R - \iota_N)\left[F_\psi\left(\frac{c_R^{Z=0}}{\iota_R-\iota_N}\right) - F_\psi\left(\frac{c_R^{Z=1}}{\iota_R-\iota_N}\right)\right] > 0$

The quantities estimated by a difference-in-means estimators on each outcome are $\Delta_{\mathcal{R}} = \rho SACE_{\mathcal{R}} > 0$ and $\Delta_{\mathcal{V}} = \rho SACE_{\mathcal{V}} > 0$.

The critical distinction between Models #1 and #2 is an assumption about the presence of post-treatment selection. Without such selection, the ATEs are identified; with such selection, the ATEs are undefined and therefore unidentified, despite the fact that the experiment remains identical.

Case #3: Endogenous Crime

Now suppose that crime may be endogenous to the see something say something campaign. The suspect commits crime when the benefit from committing crime exceeds the expected cost of getting caught. In this case, the campaign affects reporting through two channels. Conditional on a crime occurring, the lower net cost of reporting in treatment enlarges the set of bystanders (values of ψ) that would report the crime. However, this also changes the suspect's calculus. They are less likely to commit the crime if they are more likely to be reported. These effects are countervailing: treatment reduces crime rates (where there is no reporting) but increases reporting conditional on crime occurrence.

As in Case #2, selection into crime renders both ATEs undefined, which follows directly from the game tree. The SACEs here measure differences in reporting and recording among precincts where crime would have happened regardless of treatment assignment. This is characterized as a threshold in λ, denoted $\tilde{\lambda}$, at which the suspect is indifferent between committing the crime and not committing the crime when $Z = 1$.[4] While the SACE may be different from Case #2 depending on the joint distribution of λ and ψ, it is positive. This occurs because the SACE estimands effectively "close off" the crime (selection) channel.

Remark 3. *When crime occurs endogenously, then:*

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

The quantities estimated by a difference-in-means estimator on each outcome are:

$$\Delta_{\mathcal{R}} = SACE_{\mathcal{R}} - (F_{\lambda}(\tilde{\lambda}) - F_{\lambda}(\underset{\sim}{\lambda}))F_{\psi}\left(\frac{c_R^{Z=0}}{\iota_R - \iota_N} \mid \lambda \in (\underset{\sim}{\lambda}, \tilde{\lambda}]\right)$$

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

[4] As a result, the always-survivor stratum is defined by $\lambda > \tilde{\lambda}$.

Both expressions are ambiguous in sign.

However, Remark 3 shows that a naive difference-in-means estimate, does not recover $SACE_{\mathcal{R}}$ or $SACE_{\mathcal{V}}$. The ambiguous sign of these estimates reflects the countervailing channels through which the "see something say something" campaign can influence reporting and, in turn, investigation. While the identification challenges are the same across Cases #2 and #3, the interpretation challenges are different since the bystander is only strategic in Case #3. As a result of this change in our theoretical assumptions, endogenous post-treatment selection into crime renders the estimates $\Delta_{\mathcal{R}}$ and $\Delta_{\mathcal{V}}$ incapable of falsifying any theoretical predictions.

Case 4: Strategic Officer

In a final case that is closely tied to Case #3, crime remains endogenous and the officer is treated as a strategic actor. While the equilibrium reflects the fact that the officer's reporting decision is strategic, the equilibrium remains substantively similar. In equilibrium, police investigate reported cases with higher probability than non-reported cases. Note that the thresholds $\underset{\sim}{\lambda}$ and $\tilde{\lambda}$ may be slightly different from the previous case given different rates of investigation.

Remark 4. *When crime occurs endogenously and the officer is strategic, then:*

1. *$ATE_{\mathcal{R}}$ and $ATE_{\mathcal{V}}$ are undefined.*
2. $SACE_{\mathcal{R}} = F_{\psi}\left(\frac{c_R^{Z=0}}{F_{\kappa}(\alpha_{\neg r})-F_{\kappa}(\alpha_r)} \mid \lambda > \tilde{\lambda}\right) - F_{\psi}\left(\frac{c_R^{Z=1}}{F_{\kappa}(\alpha_{\neg r})-F_{\kappa}(\alpha_r)} \mid \lambda > \tilde{\lambda}\right) > 0$
   $SACE_{\mathcal{V}} = (F_{\kappa}(\alpha_{\neg r}) - F_{\kappa}(\alpha_r))\left[F_{\psi}\left(\frac{c_R^{Z=0}}{F_{\kappa}(\alpha_{\neg r})-F_{\kappa}(\alpha_r)} \mid \lambda > \tilde{\lambda}\right) - F_{\psi}\left(\frac{c_R^{Z=1}}{F_{\kappa}(\alpha_{\neg r})-F_{\kappa}(\alpha_r)} \mid \lambda > \tilde{\lambda}\right)\right] >$
   0

The quantities estimated by a difference-in-means estimator on each outcome are:

$$\begin{aligned}
\Delta_{\mathcal{R}} =& SACE_{\mathcal{R}} - (F_{\lambda}(\tilde{\lambda}) - F_{\lambda}(\underset{\sim}{\lambda}))F_{\psi}\left(\frac{c_R^{Z=0}}{F_{\kappa}(\alpha_{\neg r}) - F_{\kappa}(\alpha_r)} \mid \lambda \in (\underset{\sim}{\lambda}, < \tilde{\lambda}]\right) \\
\Delta_{\mathcal{V}} =& SACE_{\mathcal{V}} - (F_{\lambda}(\tilde{\lambda}) - F_{\lambda}(\underset{\sim}{\lambda}))\left[(F_{\kappa}(\alpha_{\neg r}) - F_{\kappa}(\alpha_r))F_{\psi}\left(\frac{c_R^{Z=0}}{F_{\kappa}(\alpha_{\neg r}) - F_{\kappa}(\alpha_r)} \mid \lambda \in (\underset{\sim}{\lambda}, \tilde{\lambda}]\right)\right]
\end{aligned}$$

Both expressions are ambiguous in sign.

As in Cases #2-#3 where there exists some form of selection into crime, the relevant ATEs are undefined. As in Case #3, the quantity estimated by a difference-in-means estimator, is ambiguously signed. The purpose of discussing this case is to demonstrate that adding a strategic actor *after* post-treatment selection does not impact (lack of) identification of the ATEs. Changes in interpretation of the SACE from Case #3 to #4 are subtle.

5 Implications for Research Design

5.1 Undefined Potential Outcomes in Social Science

Social science contains many research designs with some form of post-treatment selection, and therefore phantom counterfactuals. This selection undermines causal identification of the ATE on outcomes realized after selection. Table 3 lists examples of post-treatment selection across subfields in political science. These examples all seek to make causal claims, as elaborated in the second column. They do so, in general, by estimating some form of ATE or average treatment effect on the treated (ATT).[5]

For each of the cited literatures, I describe the potential post-treatment selection in the third column. While the selection mechanism varies by literature, there are several important commonalities. First, as in clinical studies, selection sometimes occurs through death, as in the conflict literature. Whereas individuals who survive conflict continue to act after the conflict, those who perish cannot, rendering those potential outcomes undefined. Similarly, in longer-run causal studies, selection can similarly occur through birth. For example, Hall, Huff, and Kuriwaki (2019) show that winners of an 1832 land lottery in Georgia had more sons than losers of the land lottery. We can measure the actions for individuals who are born, but the actions of individuals who would have been born under a different treatment assignment are undefined. Importantly, long-run historical studies show that this type of survival can also occur at the *cluster* level when historical communities do not persist.

Second, in social science, post-treatment selection can be much subtler than birth or death.

[5] Proposition A3 (p. A-5) generalizes the result from Proposition 1 to the ATT.

	Literature/Example	Treatment	Outcome	Post-treatment selection
1	Effects of conflict (Blattman, 2009)	Individual or community exposure to conflict.	Individuals' political attitudes or behaviors.	Death during conflict.
2	Downstream effects of shocks on political behavior. (Hall, Huff, and Kuriwaki, 2019)	Shock (i.e., wealth shock)	Descendants' political behavior.	Different descendant populations (i.e., different rates of reproduction).
3	Long-run effects of historical institutions on current outcomes (Jha, 2013)	Imposition of (pre-)colonial institutions in (pre-)colonial-era communities	Individual or community-level economic/political outcomes in present communities	Community non-persistence from (pre-)colonial era to present, different patterns of individual survival, different patterns of marriage and reproduction.
4	Email audit experiments (White, Nathan, and Faller, 2015)	Petitioner/petition characteristics	Quality of response (accuracy, respect etc.)	Subject does not respond to email.
5	Ideological positioning (Adams, 2012)	Electoral performance, t	Platform (ideology) in election $t+1$	Party ceases to exist in election $t+1$
6	Incumbency (dis)advantage (Erikson, 1971; Erikson and Titiunik, 2015)	Incumbency	Vote share of incumbent candidate or party in election $t + 1$	Candidate does not run in election $t + 1$.
7	Police use of force (Knox, Lowe, and Mummolo, 2020)	Race of citizen	Police use of force during arrest	Arrest or police contact.

Table 3: Select examples of the "truncation by death" problem across subfields and research designs in political science.

For example, in email-based audit experiments, we often conceptualize a politician or bureaucrat's response to a citizen query as a decision to respond and then a subsequent determination of content. A decision not to respond renders the potential outcomes measuring response content undefined. In other literature, researchers estimate effects of a treatment (i.e., incumbency) on voters' choices in the subsequent election. However, candidates or parties decide whether they are running before voters vote. When an incumbent, for example, chooses not to run, a voter cannot vote for the incumbent in election $t + 1$. Finally, Knox, Lowe, and Mummolo (2020) articulate post-treatment selection as a threat to inference about police use of force. In this context, police officers observe citizen race, then decide whether to stop the citizen before they face the decision of whether to use force. As such, when a civilian is not stopped, potential outcomes measuring use of force are undefined.

These examples show that phantom counterfactuals should be a cause for concern in diverse literatures in the social sciences. The class of causal research designs that I emphasize, those with sequential behavioral outcomes, is clearly quite large. Beyond standard empirical identification assumptions – like parallel trends in difference-in-difference designs – I show that theoretical assumptions about sequence of actors and their available strategies are necessary to clarify which estimands are identified in these literatures. As in the stylized examples, different theoretical assumptions imply the identification of different estimands.

5.2 Guidance for Research Design

The discussion to this point focuses on the challenges of the identification and interpretation of causal estimands in settings with sequential post-treatment behavior. Here, I turn to discussion of how considerations of possible phantom counterfactuals should inform empirical research designs. When researchers study multiple behavioral outcomes, they should assess the potential for phantom counterfactuals by specifying at least a minimal theory describing the sequence of actions and available strategies. Such a theory is important for diagnosing the problem and for determining which research design strategies may successfully mitigate issues of phantom counterfactuals.

I propose three classes of strategies in Table 4. These approaches have been used to varying degrees in existing applications, as shown by the citations in the table. However, these approaches have not yet been organized as a response to a common identification problem caused by post-treatment selection. These design recommendations are organized in three panels, focused on whether the solution emphasizes the estimation strategy given a treatment and outcome (Panel A), changes in how a treatment is assigned (Panel B), or changes in how outcomes are defined or measured (Panel C). It is important to have multiple approaches to remedying phantom counterfactuals because the feasibility of each strategy varies across applications. In discussing these strategies, I consider the scope for application of this guidance in experimental versus observational research designs, as well as in existing versus prospective studies.

Strategy #1: Estimate only defined and identified estimands: Most obviously, the results of this paper suggest that estimand identification follows from a (minimal) theory. Researchers can avoid identification issues caused by phantom counterfactuals by postulating a theory and estimating the causal quantities that are identified under their theory. Panel A of Table 4 articulates this guidance. In this approach, a researcher need not change the treatment or outcomes. Instead, they can focus on estimating the relevant identified estimands before and after the first strategy set asymmetric history. This follows closely from the policing example. The primary benefit of this strategy is that it does not require researchers to collect any additional data or reassign treatment. It can be pursued in both experimental and observational studies, as well as in both prospective and existing applications, including all of the studies cited in Table 3. This straightforward guidance departs from current practice because few researchers presently estimate SACEs in the social sciences.

This practical guidance should be tempered by the limits of what can be learned from SACEs. From a practical perspective, given estimation challenges, the SACE is often reported as an interval estimate produced by bounding estimators or sensitivity analysis (Zhang and Rubin, 2003; Aronow, Baron, and Pinson, 2019; Knox, Lowe, and Mummolo, 2020). These bounds can be quite wide, making predictions about behavior harder to falsify. Further, using an SACE to inform policy or normative discussions may be quite limited given the purposeful emphasis on a "partial

	Recommendation	Description in reference to asymmetric strategy set in Figure 1	Example
PANEL A: ESTIMATE ONLY DEFINED AND IDENTIFIED ESTIMANDS			
1	Estimate the ATE (etc.) only until the first history with an asymmetric strategy set.	Estimate $ATE = E[A\|Z = 1] - E[A\|Z = 0]$ but abstain from estimating causal effects on B. Accordingly, it may be unnecessary to measure behavior at the second history (player 2's action).	Coppock (2019)
2	Estimate the SACE (using a point or interval estimator) after the first history with an asymmetric strategy set.	Estimate player 2's choice ($B \in \{b, \neg b\}$) among "always survivor" interactions in which player 1 would $A = a$ for any Z. Formally, $SACE = E[B\|Z = 1, A = a] - E[B\|Z = 0, A = a]$.	Knox, Lowe, and Mummolo (2020)
PANEL B: RE-RANDOMIZE AT HISTORIES WHERE SELECTION OCCURS			
3	Add ancillary experiment(s) at histories with asymmetric strategy sets.	Re-randomize treatment (or some variant thereof) at the second history for all interactions in which player 1 chooses $A = a$ and estimate the ATE (etc.) for each experiment.	Golden, Gulzar, and Sonnet (2019)
PANEL C: CHANGE THE SET OF OUTCOMES			
4	Measure more outcomes prior to the first strategy set asymmetric history.	Measure additional outcomes that occur prior to or contemporaneously with A.	Slough (2020)
5	Redefine potential outcomes to reduce the threat of selection.	Redefine outcomes such that the strategy set asymmetric game can be conveyed as strategy set symmetric.	Erikson (1971), Erikson and Titiunik (2015)
6	Flatten a sequence of actions into a categorical outcome.	Collapse over the first two histories to define interaction-level outcomes, i.e. $ATE = E[a \cap b\|Z = 1] - E[a \cap b\|Z = 0]$. (This is more common in settings with a single actor.)	Findley, Nielson, and Sharman (2014)

Table 4: Design recommendations. These recommendations refer to the left panel in Figure 1 (the asymmetric strategy set). $A \in \{a, \neg a\}$ refers to the measured outcome at the first history and $B \in \{b, \neg b\}$ refers to the measured outcome at the second history.

equilibrium" effect which purges effects of selection (Joffe, 2011).

Existing applications of the SACE typically occur outside strategic settings, emphasizing sequential decisions by a single actor. For example, in Knox, Lowe, and Mummolo (2020) a police officer makes contact with a citizen before deciding whether to use force. Interval estimates of the SACE from the aforementioned bounding estimators on post-selection outcomes are expressed as a function of the previous outcome (i.e., rates of police contact). Yet, in the strategic setting depicted in the left panel of Figure 1, Player 2's determination of b or $\neg b$ does not depend on Player 1's decision, conditional on arriving at history $H^1 = a$. This follows directly from the logic of backward induction. As such, in strategic contexts, interval estimates of the SACE do not allow researchers to fully isolate the effect of treatment of Player 2's actions.

Strategy #2: Re-randomize at histories where selection occurs: Panel B of Table 4 suggests using multilevel random assignment to study sequential interactions. In this case, researchers would randomize a different treatment, say Z', at the first strategy set asymmetric history. In the left panel of Figure 1, this re-randomization of Z' would occur at the history $H^1 = a$. With this ancillary experiment, researchers could identify the ATE of Z' on Player 2's determination of b or $\neg b$ when the ATE of Z on b and $\neg b$ is undefined. This approach is advocated by Green and Tusicisny (2012) in the context of lab experiments. Golden, Gulzar, and Sonnet (2019) exemplify this approach in a field setting. Re-randomization permits identification of a distinct ATE subsequent to post-treatment selection.

Unlike Strategy #1, Strategy #2 is feasible in a much smaller set of research designs. Indeed, re-randomization is seemingly infeasible in 6 of the 7 literatures in Table 3 (see Table A4, p. A-15). Re-randomization requires researchers to manipulate a treatment at a specific point (history) in a causal process. This means that re-randomization is generally possible only for contemporaneous studies, since we cannot time travel to add an ancillary treatment in the past. This limits our ability to re-randomize in existing studies. Manipulating a second treatment may also be infeasible or unethical for other reasons, further delimiting the scope of this strategy.

Less obviously, even when it is feasible, re-randomization may not identify theory- or policy-relevant ATEs. Ancillary experiments permit the identification of cleaner "partial equilibrium" effects. Nevertheless, the identification of multiple "partial equilibrium" effects of related – but distinct – treatments does not necessarily provide any insight into the (general) equilibrium of a model. This weakness parallels critiques of the SACE as capturing only a partial equilibrium effect (Joffe, 2011). The primary benefit of re-randomization over simply estimating the SACE (like Strategy #1) is thus the identification of an ATE, which is easier to estimate than an SACE.

Theory should play an important role in the design of experiments with multilevel random assignment. Specifically, if the goal of an ancillary treatment is to identify a related ATE in the presence of post-treatment selection, it is important to have a model of when post-treatment selection occurs. By identifying histories with asymmetric strategy sets, researchers can determine when an ancillary treatment is analytically useful.

Strategy #3: Change the set of outcomes: Panel C suggests three changes in the measurement of outcome variables may help to address phantom counterfactuals. These strategies are generally simpler or cheaper to implement than ancillary experimentation. Researchers can often collect ancillary variables measuring historical causal processes. This broadens the scope of this strategy in both existing studies and original work focused on past events.

First, in clinical settings of "truncation by death," researchers often search for clinical markers that present quickly, ideally prior to death (selection). In the social science setting, researchers may gain leverage by measuring additional outcomes that present prior to the first non-strategy set symmetric history. These outcomes may provide additional leverage to validate a theory's assumptions or evaluate additional implications.

Second, researchers may redefine outcomes to reduce the threat of post-treatment selection. Returning to the incumbency advantage example, one could move from defining incumbency at the *candidate* level to defining incumbency at the *party* level (Fowler and Hall, 2014). In contexts like the US in which competitive elections generally draw candidates from both major parties, the threat that a party will not run a candidate is minimal. This is akin to ensuring that the challenger (party) always contests election $t + 1$. Note that this suggestion departs from calls to estimate incumbency advantage unconditional on running by imputing a “0” when a candidate does not run as advocated by De Magalhaes (2017).

Finally, researchers may “flatten” a sequence of outcomes into a categorical measure. For example, Findley, Nielson, and Sharman (2014) study responses of agents of business incorporation services to “mystery shopper” email requests for incorporation with experimental manipulations. They “flatten” the agent’s sequential decision of (1) whether to respond; and (2) the content of response into a categorical measure including non-response and each type of content. This strategy precludes the content potential outcomes from being undefined in the case of non-response. One requirement for the ability to “flatten” sequential outcomes is that the flattened outcomes are measured. In the policing example, for example, it may be interesting to decompose precincts reporting crime, precincts with unreported crime, and no-crime precincts. However, crime is latent in the example. This limits our ability to distinguish precincts with unreported crime from those with no crime, limiting our ability to flatten these outcomes.[6]

Changes in the measurement or definition of outcomes can permit identification of the ATE or ATT for outcomes with phantom counterfactuals. This strategy is attractive because it feasible in observational and experimental studies and, in some cases, with existing data. However, when we redefine outcomes, we may introduce new challenges for interpretation. When redefinition leads to non-obvious outcome measures, theory can help to provide clearer predictions for these outcome measures, in addition to justifying the use of these redefined outcomes in service of identification.

Like most research design guidance, the design recommendations in Table 4 each have strengths and weaknesses. I summarize the scope for application of these methods along with these pros and cons of each method in Table A3 (p. A-14). Further, in Table A4 (p. A-15), I illustrate these recommendations in the context of each of the literatures in Table 3.

[6] Of course, latent crime incidence could be measured via crime victimization surveys or on-the-ground audits.

5.3 Identification of Treatment Effects on Equilibrium Outcomes

To this point, I have focused on potential outcomes measuring the behavior of individual actors. When treatment is assigned at the level of a strategic interaction, we can measure the causal effects of treatment on equilibrium outcomes created by the actions of multiple actors.[7] By assignment at the interaction level, I refer to clustered assignment schemes in which all interactions occur within clusters. In this setting, ATEs and ATEs on equilibrium outcomes are identified even when potential outcomes measuring some subject's behavior are undefined.

Consider, for example, a two-player public goods game. First, Player 1 decides whether to contribute to the public good. If and only if she fails to contribute, then Player 2 chooses whether to contribute or not. If one player contributes, the public good is provided. This game corresponds to the left panel of Figure 1. If a randomly-assigned treatment increases the value of the public good, we may be interested in measuring rates of contribution. Per the previous discussion of Figure 1, the ATE on Player 1's contribution is identified but the ATE on Player 2's contribution is not, because he faces no decision to contribute if Player 1 contributed. However, we could measure instead whether the public good is provided – an outcome which is jointly determined by the actions of Player 1 and Player 2. Following the analogy from the previous section, this outcome effectively flattens both players' behavior into a single equilibrium outcome, for which both counterfactuals are defined.

More generally, researchers can identify ATEs on equilibrium outcomes or equilibrium selection even when one they cannot identify the ATE on all actors' behavior. In these designs, the focus is generally not the behavior of any single actor, but manifestations of some interaction. Using theory to characterize an equilibrium (or equilibria) is therefore critically important to the interpretation of flattened equilibrium outcomes in strategic settings.

[7] The natural analogue to equilibrium outcomes in decision theoretic settings are the "flattened" sequential outcomes discussed in the previous section.

6 Conclusion

This paper highlights limits to causal identification in studies with multiple, possibly sequential, behavioral outcomes. In these settings, post-treatment selection generates phantom counterfactuals, or undefined potential outcomes for subsequent outcomes. Phantom counterfactuals undermine the identification of the ATE or ATT when present. In this class of research designs, therefore, these estimands are identified by a research design for specific outcomes. Applied theory should guides our determination of these outcomes, and thus our choice of estimands.

I show that phantom counterfactuals are widely present in social science research. Failure to identify phantom counterfactuals undermines claims to causal identification. As such, phantom counterfactuals should be addressed through careful research design. I provide three courses of action for applied scholars who confront phantom counterfactuals in their work. These strategies have been employed in existing literature but have not been previously synthesized as a response to a common threat to identification.

The ultimate insights of this paper provide guidance on how applied theory can support claims to identification and the design of empirical studies. Separating applied theory from research design limits our ability to make inferences about data in a variety of common settings in social science. Theory can guide researchers' choice of outcomes and the estimation strategy employed to strengthen the credibility of claims to causal identification. Ultimately, this paper calls for a more explicit marriage of theory and data in identification-oriented empirical work.

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.

Aronow, Peter M., Jonathon Baron, and Lauren Pinson. 2019. "A Note on Dropping Experimental Subjects who Fail a Manipulation Check." *Political Analysis* Forthcoming.

Ashworth, Scott, Christopher R. Berry, and Ethan Bueno de Mesquita. 2015. "All Else Equal in Theory and Data (Big or Small)." *PS Political Science* 48 (1): 89–94.

Ashworth, Scott, and Ethan Bueno de Mesquita. 2014. "Is Voter Competence Good for Voters? Information, Rationality, and Democratic Performance." *American Political Science Review* 565-587.

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

Bueno de Mesquita, Ethan, and Scott A. Tyson. 2020. "The Commensurability Problem: Conceptual Difficulties in Estimating the Effect of Behavior on Behavior." *American Political Science Review* 2 (375-391).

Clark, William Roberts, and Matt Golder. 2015. "Big Data, Causal Inference, and Formal Theory: Contradictory Trends in Political Science?" *PS Political Science* 48 (1): 65–70.

Coppock, Alexander. 2019. "Avoiding Post-Treatment Bias in Audit Experiments." *Journal of Experimental Political Science* 6 (1): 1–14.

Coppock, Alexander, Alan S. Gerber, Donald P. Green, and Holger L. Kern. 2017. "Combining Double Sampling and Bounds to Address Nonignorable Missing Outcomes in Randomized Experiments." *Political Analysis* 25: 188–206.

De Magalhaes, Leandro. 2017. "Incumbency Effects in a Comparative Perspective: Evidence from Brazilian Mayoral Elections." *Political Analysis* 23 (1): 113–126.

Eggers, Andrew. 2017. "Quality-Based Explanations of Incumbency Effects." *Journal of Politics* 79 (4): 1315–1328.

Erikson, Robert S. 1971. "The Advantage of Incumbency in Congressional Elections." *Polity* 3 (3): 395–405.

Erikson, Robert S., and Rocio Titiunik. 2015. "Using Regression Discontinuity to Uncover the Personal Incumbency Advantage." *Quarterly Journal of Political Science* 10: 101–119.

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.

Fowler, Anthony, and Andrew B. Hall. 2014. “Disentangling the Persnal and Partisan Incumbency Advantages: Evidence from Close Elections and Term Limits.” *Quarterly Journal of Political Science* 9: 501–531.

Frangakis, Constantine E., and Donald B. Rubin. 2002. “Principal Stratification in Causal Inference.” *Biometrics* 58 (1): 21–29.

Franzese, Robert. 2020. *The SAGE Handbook of Research Methods in Political Science and International Relations*. London: SAGE Publications chapter Econometric Modeling: From Measurement, Prediction, and Causal Inference to Causal-Response Estimation, pp. 577–598.

Gailmard, Sean. 2021. “Theory, History, and Political Economy.” *Journal of Historical Political Economy* 1: 69–104.

Gailmard, Sean, and John W. Patty. 2018. “Preventing Prevention.” *American Journal of Political Science* 63 (2): 342–352.

Golden, Miriam, Saad Gulzar, and Luke Sonnet. 2019. ““Press 1 for Roads”: Motivating Programmatic Politics in Pakistan.” Working paper.

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

Green, Donald P, and Andrej Tusicisny. 2012. “Statistical Analysis of Results from Laboratory Studies in Experimental Economics: A Critique of Current Practice.” Available at `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2181654`.

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.

Heckman, James J. 2008. “Econometric Causality.” *International Statistical Review* 76 (1): 1–27.

Holland, Paul W. 1986. “Statistics and Causal Inference.” *Journal of the American Statistical Association* 81 (396): 945–960.

Huber, John D. 2017. *Exclusion by Elections: Inequality, Ethnic Identity, and Democracy*. New York: Cambridge University Press.

Hume, David. 1739-40 (2003). *A Treatise of Human Nature*. Penguin Books.

Izzo, Federica, Torun Dewan, and Stephane Wolton. 2020. “Cumulative Knowledge in the Social Sciences: The Case of Improving Voters’ Information.” Working Paper available at `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3239047`.

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

Joffe, Marshall. 2011. “Principal Stratification and Attribution Prohibition: Good Ideas Taken Too Far.” *International Journal of Biostatistics* 7 (1): 35.

Kant, Immanuel. 1781 (1996). *Critique of Pure Reason*. Indianapolis: Hackett Publishing.

Keane, Michael P. 2010. “Structural vs. atheoretic approaches to econometrics.” *Journal of Econometrics* 156: 3–20.

King, Gary, James Honaker, Anne Joseph, and Kenneth Scheve. 2001. “Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation.” *American Political Science Review* 95 (1): 49–69.

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

McConnell, Sheena, Elizabeth A. Stuart, and Barbara Devaney. 2008. “The Truncation-by-Death Problem: What to do in an Experimental Evaluation When the Outcome is Not Always Defined.” *Evaluation Review* 32 (2): 157–186.

Montgomery, Jacob M., Brendan Nyhan, and Michelle Torres. 2018. “How Conditioning on Post-treatment Variables Can Ruin your Experiment and What to Do about It.” *American Journal of Political Science* 62 (3): 760–775.

Prato, Carlo, and Stephane Wolton. 2019. “Electoral Imbalances and their Consequences.” *Journal of Politics* First View: 1–15.

Rubin, Donald B. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons Inc.

Rust, John. 2010. “Comments on “Structural vs. atheoretic approaches to econometrics” by Michael Keane.” *Journal of Econometrics* 156: 21–24.

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

Shadish, William R., Thomas D. Cook, and Donald T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston, MA: Houghton Mifflin.

Signorino, Curtis S. 2003. “Structure and Uncertainty in Discrete Choice Models.” *Political Analysis* 11: 316–344.

Signorino, Curtis S., and Kuzey Yilmaz. 2003. “Strategic Misspecification in Regression Models.” *American Journal of Political Science* 47 (3): 551–566.

Slough, Tara. 2020. “Bureaucrats Driving Inequality in Access: Experimental Evidence from Colombia.” Working paper available at `http://taraslough.com/assets/pdf/colombia_audit.pdf`.

Slough, Tara, and Scott A. Tyson. 2021. “External Validity and Meta-Analysis.” Available at `http://taraslough.com/assets/pdf/ev_ma.pdf`.

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.