On Theory and Identification:
When and Why We Need Theory for Causal Identification

Tara Slough*

May 6, 2021

Abstract

What is the role for theory in identification-driven research designs? I argue that in a substantial class of research designs, theory is necessary for the identification of standard reduced-form causal estimands. Specifically, I show that when empiricists study a sequence of behavioral outcomes, post-treatment selection can render standard causal estimands undefined and thus unidentified, even when standard identification assumptions hold. Using a stylized example of crime, reporting, and recording, I illustrate how different articulations of a theory posit different sets of identified estimands, all while holding constant an experimental design. I generalize these observations to any dynamic model of post-treatment behavior. In so doing, I show that claims to identification of treatment effects on potentially sequential behavioral outcomes imply a set of theoretical assumptions, whether or not they are stated explicitly. This paper advocates a closer marriage of theory and empirics in identification-driven research.

*Assistant Professor, New York University, tara.slough@nyu.edu. I thank Neal Beck, Alex Coppock, Scott de Marchi, Bob Erikson, Don Green, Justin Esarey, Daniel Hidalgo, Macartan Humphreys, Thomas Leavitt, Winston Lin, Kevin Munger, Fredrik Sävje, Mike Ting, Scott Tyson, Stephane Wolton, 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. This project is supported in part by an NSF Graduate Research Fellowship, DGE-11-44155.

1 Introduction

The influence of the identification or credibility revolution on social science research raises questions about the role of applied theory in empirical research (Franzese, 2020; Clark and Golder, 2015; Ashworth, Berry, and Bueno de Mesquita, 2015; Samii, 2016; Huber, 2017). In political science, scholars debate whether there exist tensions between the goals of theory and causal identification, stated provocatively by Huber's (2013) question "is theory getting lost in the 'identification revolution?" In this paper, I argue that for a large class of identification-driven research designs, a theory of post-treatment behavior is necessary for the identification of standard reduced-form causal estimands. Focusing on studies with sequential behavioral outcomes, I characterize the conditions under which theory is necessary for identification. I find that in many empirical settings, the absence of explicit theory of the sort lamented by Huber (2013) may undermine claims to causal identification.

The identification revolution in social science emphasizes research designs that invoke fewer and/or less heroic modeling assumptions to identify causal quantities of interest (Angrist and Pischke, 2010; Aronow and Miller, 2019). The reduced set of assumptions invoked in these research designs focus on what happens *before* treatment like how treatment is assigned and how treatment assignment maps onto the treatments that are ultimately delivered. In this paper I argue that the structure of what happens *after* treatment also poses underappreciated limits to identification of causal estimands, beyond typical discussions of non-interference (SUTVA). Applied theory provides necessary assumptions to structure thinking about responses to treatment. The crucial requirement of a theory in this context is therefore that it models: (1) relationships between an exogenous treatment and endogenous outcomes; and (2) relevant relationships between endogenous outcomes.[1] I argue that such theories are necessary for the identification of standard, reduced-form causal estimands in identification-driven research when behavioral outcomes may be sequential.

My argument begins from the observation that "truncation by death" problems are ubiquitous

[1] A basic reading of the potential outcomes framework may suffice for the former, but does not typically incorporate dependencies between outcomes.

in the social sciences. The term "truncation by death" was developed in clinical studies in which participants may die between the time that treatment is assigned and ultimate outcomes of interest are measured (e.g., Zhang and Rubin, 2003; McConnell, Stuart, and Devaney, 2008). A study participant's death renders subsequent potential outcomes undefined. Because estimands such as the average treatment effect (ATE) are defined in terms of expectations evaluated over potential outcomes, an undefined potential outcome for any unit renders the ATE undefined and therefore unidentified (Holland, 1986). In this article, I formalize a link between "truncation by death" problems in empirical social science and the extensive form representation of theoretical models. This link between the structure of theoretical models and identification of standard causal estimands underscores the frequency with which "truncation by death" problems emerge in social science.

To illustrate this link between theoretical models 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 to increase rates of reporting. Researchers aim to study crime reporting with administrative 911 reports and measures of crime incidence. Through four variants of the model, I show two main findings. First, any selection into crime is analogous to death in clinical studies. It renders the ATE on subsequent outcomes (including 911 calls and administrative crime measures) undefined. Second, standard estimators employed to estimate the ATE do not recover the well-defined and identified causal effect among those neighborhoods where a crime would always occur regardless of treatment.[2] Moreover, when selection into crime is endogenous to the treatment – because the suspect anticipates higher rates of being caught – the estimates produced by standard estimators cannot falsify any theoretical prediction.

Generalizing from this illustration, the central result of this paper characterizes the outcomes for which an ATE is identified in terms of the extensive form of a theoretical model. Specifically, I show that the ATE is identified only for outcomes measuring actions prior to the first history of a model in which an actor or her set of available strategies depends on the realization of a previous action. This result has two implications for the practice of applied empirical research. First, a

[2]This estimand is most commonly called the survivor average causal effect (SACE).

claim to identify the ATE on sequential behavioral outcomes implies strong assumptions about the underlying extensive form, whether they are made explicit or not. Second, I show that in the context of potentially sequential behavioral outcomes, the ability of a research design to identify causal effects is specific to *estimand* and *outcome* pairs.

In light of these results, I provide guidance for research design. I show that researchers can mitigate these identification problems by: (1) changing the set of estimands; (2) re-randomizing treatment; or (3) changing or redefining outcomes. Critically, all of these recommendations follow from an explicit link between a minimal theoretical model and the research design. For this reason, I advocate a closer marriage of theory and identification-oriented research designs.

This work contributes to discussions of the role of theory in identification-driven empirical research. First, like work comparing structural and reduced-form approaches to causality, this paper questions the possibility of "agnostic" inferences – those made without reference to a theory (Heckman, 2008). These works have focused on the limited interpretability of reduced-form estimates in the absence of theory (Keane, 2010; Rust, 2010) and the perils of misspecification of statistical models as they relate to underlying theory (Signorino, 2003; Signorino and Yilmaz, 2003). This paper departs from these works in two ways. First, the discussion focuses on the "reduced form" causal estimands popularized by the Neyman-Rubin causal model. While one could certainly use the issues described in this paper to motivate the adoption of structural estimation, the issues I identify also provide guidance for practitioners of reduced form causal approaches, the dominant current practice. Second, I focus on whether causal estimands are defined under a theoretical model. This is distinct from the questions of interpretation and bias that animate existing discussions of the relationship between theory and empirics. Indeed, if an estimand is undefined, the bias of its estimator is similarly undefined.

My argument contributes to a new literature on the "theoretical implications of empirical models" (TIEM), an "inversion" of an established literature on the "empirical implications of theoretical models" (Morton, 1999). The most common approach to TIEM involves writing a model to interpret existing empirical findings (e.g., Ashworth and de Mesquita, 2014; Gailmard and Patty,

2018; Prato and Wolton, 2019; Izzo, Dewan, and Wolton, 2020; Sun and Tyson, 2019). A second approach examines a specific research design or class of theoretical models to examine the validity of the design on the basis of an underlying theory (e.g., Eggers, 2017). My argument adopts the second approach, emphasizing one feature of empirical studies – (potentially) sequential behavioral outcomes – that is common to many identification-driven research designs. Following Bueno de Mesquita and Tyson (2020), I articulate a class of commensurability problems, referring here to situations in which analysts aim to estimate a quantity that is theoretically undefined.[3]

The focus on what happens after treatment represents an increasing concern in research design. Yet existing works largely focus on the ills of "bad" controls (Montgomery, Nyhan, and Torres, 2018); post-treatment sample conditioning (Aronow, Baron, and Pinson, 2019); or post-treatment selection in various empirical applications (Knox, Lowe, and Mummolo, 2020; Coppock, 2019). The treatment in this article speaks to a wider class of applications. By linking "truncation by death" to standard features of dynamic models in political science, I show how explicit ties between applied theory and identification-oriented research designs expose the prevalence of identification problems induced by post-treatment selection.

2 Definition, Identification, and Interpretation of Causal Estimands

2.1 Sequential Outcomes

Researchers often study the effects of some treatment (or independent variable), Z, on more than one outcome. This paper focuses on a wide range of empirical contexts with sequential post-treatment outcomes. Figure 1 depicts two possible relationships between a randomly assigned treatment, Z, and two sequential outcomes, Y_1 and Y_2.[4] In both panels (a) and (b), Y_2 is a function of both Z and Y_1, denoted $Y_2(Z, Y_1)$, whereas Y_1 is only a function of Z, denoted $Y_1(Z)$. Unlike in panel (a), in panel (b), the node indicated with $\odot$ indicates that Y_2 is defined for only some values of Y_1. In other words, there exist some units for which the potential outcome $Y_2(Z, Y_1)$ is

[3]See also Abramson, Koçak, and Magazinnik (2021) for another recent articulation of a commensurability issue in conjoint surveys.

[4]This discussion is not specific to experiments and generalizes to much more complex models.

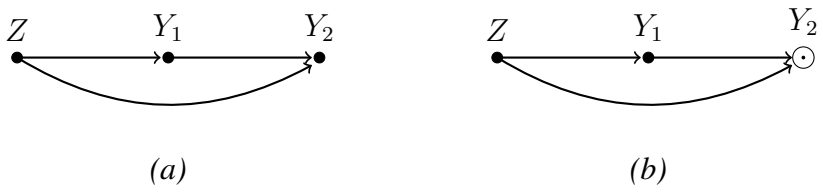


Figure 1: Two graphical causal models of a randomly-assigned treatment, Z, and two sequential outcomes, Y_1 and Y_2. The $\odot$ node indicates that outcome variable Y_2 is not defined for all levels of outcome variable Y_1.

not defined. Such undefined outcomes undermine claims of causal identification (Holland, 1986). I argue that extensions of panel (b) are very common in political science and pose underappreciated limits to causal identification.

2.2 Undefined Potential Outcomes Undermine Claims to Identification

Identification-oriented work purports to identify the causal effect of a treatment on at least one outcome. Stated more precisely, these works invoke a set of assumptions in order to identify a specific causal estimand, such as the ATE. Following Manski (1995), the process of drawing causal inferences can be separated into identification and statistical components. In this article, I focus exclusively on identification.

One requirement for identification of many causal estimands, including the ATE, is that all variables – including all potential outcomes – are defined for every unit in the experimental population (Holland, 1986). Because the ATE is defined in terms of expectations evaluated over potential outcomes, an undefined potential outcome for some unit renders these expectations, and thus the ATE, undefined. An undefined estimand is not identified.

The problem of "truncation by death" represents the best-known setting in which undefined potential outcomes arise (e.g., Zhang and Rubin, 2003; McConnell, Stuart, and Devaney, 2008). In medical studies, "truncation by death" occurs when a subject dies after treatment but prior to the measurement of the ultimate outcome of interest. For example, researchers may 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 for the quality of life

measure is undefined. Standard experimental estimators of the ATE (e.g., a difference-in-means) estimate an undefined and thus unidentified quantity. Moreover, comparison of quality of life among subjects that survive is not necessarily a principled experimental comparison because death may be endogenous to the treatment under study, undermining internal validity.

More generally, an undefined potential outcome is one in which observed and unobserved values are measured on qualitatively different scales (McConnell, Stuart, and Devaney, 2008). Death and a numeric quality of life measure, for example, exist on distinct scales. A deceased subject's quality of life is therefore undefined. The difference in scales differentiates undefined outcomes from attrition or missingness.

The distinction between undefined outcomes and attrition is clear when considering statistical methods for addressing 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.[5]

2.3 Undefined Potential Outcomes in Social Science

I contend that the social science literature is replete with research designs that parallel clinical studies with "truncation by death." A common feature of such research designs is some form of post-treatment selection prior to the realization of an outcome of interest. As in the clinical setting, in some social science settings, selection occurs by death, though this need not be the

[5]Bounding approaches on a distinct estimand, the survivor average causal effect (SACE) do resemble those used to bound interval estimates in the case of attrition, though the underlying quantity of interest is distinct.

	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 1: Select examples of the "truncation by death" problem across subfields and research designs in political science. See Table A2 for further elaboration of these examples.

case. Table 1 provides a set of examples of post-treatment selection problems akin to "truncation by death" across subfields in political science. Note that when treatment is assigned to clusters or groups of individuals, selection could occur at the cluster level (long-run development) or unit level (conflict). Importantly, aggregation of undefined individual potential outcomes cannot solve the problem described here.

Table 1 contains seven literatures that elaborate causal relationships: the effects of conflict, the downstream effects of various shocks on political behavior, the long-run effects of historic institutions, email audits, party platform positioning, incumbency advantage, and police use of force. These studies using experiments, natural experiments, regression discontinuity designs, or difference-in-difference strategies. Even if all standard identifying assumptions hold, if any potential outcomes are undefined, the general quantities of interest, typically some ATE, local average

treatment effect (LATE), or average treatment effect on the treated (ATT), are also undefined. In this sense, without the imposition of some additional structure (assumptions) on the post-treatment causal process, standard identifying assumptions may not ensure identification of these standard estimands.

One common feature of problems of "truncation by death" is that outcomes are sequential. Indeed, in the clinical setting, all experimental subjects will eventually die; quality of life outcomes are undefined if subjects die *before* realization of the quality of life measure. To this extent, the sequencing of outcomes becomes a critical assumption in understanding what estimands are identified by a research design. A second feature of the examples provided is that the selection process is behavioral, broadly speaking, as opposed to attitudinal.

When modeling a sequence of post-treatment outcomes, a fundamental concern is whether post-treatment actions alter the available set of strategies of a subsequent action. To this end, theory introduces necessary additional assumptions about the sequence and structure of multiple outcomes. For the purpose of identification, theory generates implications for what estimands could plausibly be identified. Empirically, these considerations suggest what comparisons, i.e. between treatment and control, could estimate well-defined causal quantities. Indeed, as I show by example in Section 3, different theoretical assumptions with the same research design imply the identification of different estimands. They also suggest different approaches to analysis of the data.

2.4 Alternative Estimand and Interpretation

Practitioners frequently turn to an alternative causal estimand, the survivor average causal effect (SACE) as a defined and identified estimand in the presence of "truncation by death." This is the average causal effect of a treatment among the stratum of subjects that would have survived regardless of treatment assignment. If $S(Z)$ is the potential outcome measuring post-treatment selection, here survival, researchers would ideally estimate the average causal effect for subjects for which $S(Z) = 1 \forall Z$. The causal effect of the treatment on quality of life is well-defined for this stratum as the ultimate outcome, $Y(Z)$, is defined on the same scale among survivors. For a

detailed exposition of this principal stratification approach, see Appendix A1. Unfortunately, we cannot infer membership in this stratum from the data if selection occurs because we can only observe one potential outcome for each subject, posing challenges for point estimation (Zhang and Rubin, 2003).

Estimation challenges aside, the SACE can be a useful measure for understanding why effects manifest. In effect, examining a causal effect among "always survivors," closes off selection as a causal mechanism. In the simplest case, the SACE allows for estimation of the "partial equilibrium" effect of a treatment among a sub-population, the always-survivor stratum. Yet, these comparisons can be misleading in terms of understanding broader "general equilibrium" effects which include selection (Joffe, 2011). Nevertheless, it is useful to consider the SACE as a benchmark causal estimand when the ATE is undefined.

3 Stylized Example

3.1 Why Formalize?

The primary concern of this paper is the relationship between theory and causal estimands. The mapping between theoretical predictions and reduced-form estimands is therefore central to the argument forwarded. Because estimands are expressed formally, it is useful to state the equilibrium in comparable language for purposes of illustration and derivation.

The theories enumerated here are neither complex nor counterintuitive. Yet, the mapping between theoretical predictions and relevant causal estimands is non-trivial even in these simple cases. To illustrate the identification concerns, I provide four nested theories and show the implications for analysis and interpretation of an experiment.

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

Consider a "see something, say something" campaign to increase crime reporting by citizens and crime incidence.[6] Suppose that the campaign is cluster random assigned to micro-neighborhoods

[6]This application is roughly inspired by one treatment arm of the experiment described in Arias et al. (2019).

within a city. Denote a binary treatment indicator, $Z_i \in \{0, 1\}$. Researchers measure outcomes using counts of geo-coded crime reports (911 calls or the equivalent) aggregated to the micro-neighborhood level, denoted $\mathcal{R}_i$, and geo-coded reported crime incidence data aggregated to the same level, denoted $\mathcal{V}_i$.[7]

The 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 the stable unit treatment value assumption (SUTVA) holds.[8] In standard practice, researchers would generally seek to estimate the ATE (or intent to treat effect) on reporting and crime incidence. The difference-in-means can be estimated by OLS with the specification in Equation 1 for outcomes $Y_i \in \{\mathcal{R}_i, \mathcal{V}_i\}$.

$$Y_i = \beta + \Delta Z_i + \epsilon_i \tag{1}$$

The focus of enumerating the theory revolves around whether the estimator Δ estimates the ATE or any well-defined causal estimand. I calculate the quantity estimated by Δ for each outcome for comparison to the analogous ATE and SACE. I denote these quantities $\Delta_{\mathcal{R}}$ and $\Delta_{\mathcal{V}}$, for reporting and incidence, respectively.

To preview the issues identified by the model, consider two features of this setting. First, there may exist some variation in the occurence of crime to report. Not reporting a crime that did not occur is qualitatively distinct from not reporting a crime that did occur. This distinction is a critical assumption of the models enumerated here. Second, and more specific to the empirical application, the true level of crime (or whether a crime occurred) is unobserved. In other words, the police records identify the subset of crimes that are investigated, not the set of crimes that occur.

[7] I use calligraphic lettering to denote measured outcome variables, $\mathcal{R}_i$ and $\mathcal{V}_i$. The treatment indicator Z_i is maintained in both the model and the data.

[8] General equilibrium effects are often invoked as a violation of SUTVA. This is not necessarily the case. The clustered assignment in the present design is consistent with SUTVA under all models specified here.

3.3 Four Cases of a Model

I enumerate four cases of a simple, stylized model that convey 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 parameterization of utility functions. Collectively, these assumptions ensure comparability across both game theoretic and decision theoretic models. Further, among the game theoretic models, these assumptions allow for invocation of a common equilibrium concept.

The cases each assume some subset of three players: a bystander, a suspect, and an officer, denoted B, S, and O, respectively. S decides whether or not to commit a crime, denoted v or $\neg v$. By commiting a crime, the suspect receives some surplus, $\lambda \geq 0$, drawn from the density $f_\lambda(\cdot)$ with cdf $F_\lambda(\cdot)$. However, if a suspect that commits the crime is investigated, she pays a penalty $p > 0$.

B observes whether a crime occurs. If the crime occurs, he chooses whether or not to report, at net cost $c_r > 0$. The "see something say something" campaign corresponds to a reduction in net costs of reporting, such that $c_r^{Z=1} < c_r^{Z=0}$. In principle, the campaign provides information and appeals 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 $F_\psi(\cdot)$. Importantly, I make no assumptions about properties of the joint distribution of λ and ψ.

O observes that a crime occurred and whether or not it was recorded. They choose to investigate or not to investigate. An investigation requires some effort by the officer at cost κ. κ is a random variable drawn from pdf $f_\kappa(\cdot)$, with cdf $F_\kappa(\cdot)$ and support on $[0, \bar{\kappa}]$. The officer faces the possibility of sanction, $s > \bar{\kappa}$ for failing to respond to crimes detected by a random audit. Denote the expectation of a sanction for an audited officer, e.g. s times the probability of sanction as α. Assume that the officer is audited at a higher probabiltiy for reported crimes due to increased legibility such that: $0 < \alpha_{\neg r} < \alpha_r < s$.

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 or not to report the crime.	(2) The bystander observes whether a crime was commited. If it was commited, she decides whether or not 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 commited. If it was commited, she decides whether or not to report the crime.	(2) The bystander observes whether a crime was commited. If it was commited, she decides whether or not 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.

The four cases of this model vary in their assumptions about which players are strategic. In all cases, the bystander decides whether or not 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. As is clear in Figure 2, no reporting and no investigation occur if a crime has not occurred. This has two implications for the outcomes of interest. It implies that reports comprise a subset of crimes that occur. There are no reports when the suspect (resp. nature) does not commit a crime. Second, in terms of police investigations, there are no false positives (investigations where no crimes occur). These assumptions may be too strong, but they simplify exposition in what follows.

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. The equilbrium characterizations and proofs thereof are straightforward from inspection of Figure 2 and comparison of expected utilities, and is thus relegated to Appendix A3.

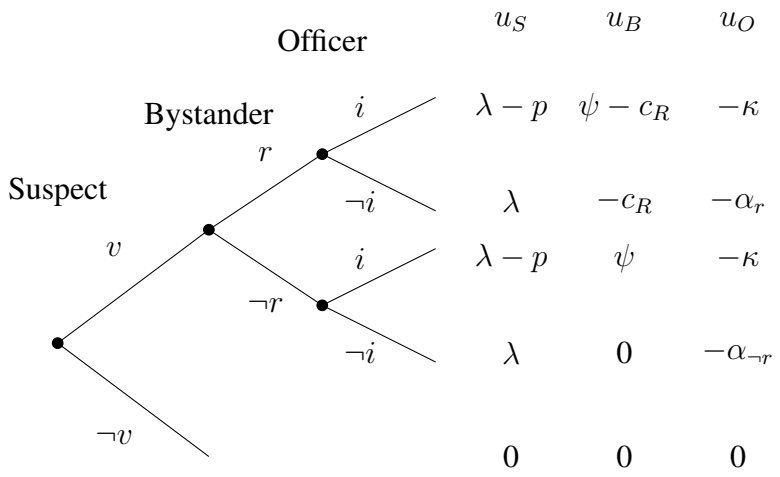


Figure 2: Extensive form representation for Case #4.

Moving from equilibrium characterizations to causal estimands requires two additional considerations. 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, i.e. $\mathcal{R}_i = 1$ if $v \cap r$ and that a case enters police records if it is investigated by police, i.e. $\mathcal{V}_i = 1$ if $v \cap i$. Second, estimands are expressed in terms of expectations 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 these outcomes in the aggregate (i.e., across precincts) 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. Here, we are only concerned with the bystander's decision of whether to report or not. As shown in Appendix A3, the bystander will report if the cost of reporting is sufficiently low relative to expected utility from the resolution of order by the police. The ATE on reporting, then, is simply the difference in proportion of bystanders reporting the crime in treatment versus control. This quantity is positive since the net costs of reporting are lower in treatment than in control. Higher levels of reporting with no change in crime occurence imply that the ATE on the recording of crime must also be positive. Because there is no selection into crime, the SACE and ATE must be equivalent.

In this case, under the standard "empirical" assumptions above, the difference-in-means estimators are an unbiased estimators of each ATE, respectively.

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 estimators on each outcome are: $\Delta_{\mathcal{R}} = ATE_{\mathcal{R}}$ *and* $\Delta_{\mathcal{V}} = ATE_{\mathcal{V}}$*. (All proofs in appendix.)*

Case #2: Exogenous Crime

Case #2 parallels Case #1 except there exists exogenous selection into crime. With probability $\rho \in (0, 1)$ a crime occurs, regardless of treatment assignment of the precinct. Because there are precincts with no crime, the bystander no longer faces the decision of whether or not to report when crime did not occur. As a result, $ATE_{\mathcal{R}}$ and $ATE_{\mathcal{V}}$ are no longer defined. In contrast, the relevant SACE estimands reflect the difference in rates of reporting and reporting among precincts in which a crime would occur regardless of treatment assignment. Because crime is exogenous, these precincts represent a random sample of all precincts. Thus, the SACEs are equivalent to the ATEs in Case #1.

However, even with *exogenous* selection, a naive difference-in-means no longer estimates the SACE. Since we do not observe true crime levels, the naive estimator effectively imputes an outcome of no reporting ($\neg r$) when crime does not occur. This equates non-reporting of crime that occurs with not reporting a crime that did not occur. Since crime is is exogeneous, however, this estimator estimates the SACE scaled by the crime rate, ρ. With the present research design and the data described here, ρ is not identifiable. Importantly, however, the difference-in-mean 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 neither defined nor identified, despite the fact that the experiment remains identical. These examples show that holding the research design constant, our theoretical assumptions posit implications for identification.

Case #3: Endogenous Crime

Now suppose that crime may be endogenous to the see something say something campaign. Crime is committed when the surplus from committing the crime exceeds the expected disutility 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. She is less likely to commit the crime if she is more likely to be reported. These effects are counterveiling: treatment reduces crime rates (where there is no reporting) but increases reporting conditional on crime occurrence. Without further assumptions on f_λ or f_ψ, it is impossible to sign the resultant difference-in-means estimates.

As in Case #2, selection into crime renders both ATEs undefined. The SACEs here measure differences in reporting 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 $Z = 1$. Within the principal

strata often used in the exposition of the SACE in Table A1, $\lambda > \tilde{\lambda}$ corresponds to an "always survivor" precinct, a place where crime always occurs, regardless of treatment assignment. Define the value of λ at which the suspect is indifferent between committing the crime and not committing the crime when $Z = 0$ as $\underset{\sim}{\lambda}$. Because treatment increases the rate of reporting, $\underset{\sim}{\lambda} \leq \tilde{\lambda}$. The interval $\lambda \in (\underset{\sim}{\lambda}, \tilde{\lambda}]$ defines the stratum of "if untreated survivor" precincts. Finally any precinct in which $\lambda < \underset{\sim}{\lambda}$ represents a "never survivor" – a precinct where crime never occurs, regardless of treatment assignment. 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:

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

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 counterveiling 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 endogenous post-treatment selection renders the estimands $\Delta_{\mathcal{R}}$ and $\Delta_{\mathcal{V}}$ incapable of falsifying any theoretical predictions. To the extent that endogenous selection into crime is plausible, the experiment does

not provide empirical leverage to identify any standard causal estimand on reporting or investigation.

The *extensive form* and the *equilibrium* of the model play two distinct roles in generating these insights. The asymmetry in the bystander's strategy sets in the extensive form (even absent utilities) indicates that the ATE will be undefined and thus unidentified. This observation does not require specification of utilities or an equilibrium characterization. However, the point on falsifiability relies on the SPNE characterized in Appendix A3. In this regard, the extensive form is critical for identification; the equilibrium is useful for interpretation.

Note that the structure of this case (but obviously not the model) parallels the structure of the Knox, Lowe, and Mummolo's 2020 account of racial bias in police use of force. In that work, the authors derive sharp nonparameteric bounds on the SACE of race on police use of force.[9] The contribution of the present exposition is to generalize this setting while drawing the parallel to an extensive form representation of behavior. If the SACE is an appropriate test of an argument, the estimator developed in Knox, Lowe, and Mummolo (2020) may find more widespread application beyond the setting of race and policing.

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 parameterization of the equilibrium reflects the fact that the officer's reporting decision is strategic, the equilibrium remains substantively equivalent. In equilibrium, police investigate reported cases with higher probability than non-reported cases. As such, the exogenous probabilities of investigation in the previous case approximate the officer's equilibrium strategy. Note that the thresholds $\underset{\sim}{\lambda}$ and $\tilde{\lambda}$ may be slightly different from the previous case given different rates of investigation, though their substantive interpretation and mapping to the strata in Table A1 is identical.

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

[9]The authors refer to the SACE as the ATE among the subset of citizens that are stopped by police – those for which the the second strategy set is defined. These estimands are equivalent.

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:

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

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

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. The SACEs are both positive and reflect only the effect of increased reporting by the bystander, as opposed to differences in rates of crime. However, the quantity estimated by a difference-in-means estimate, as in Case #3, is ambiguously signed. The purpose of discussing this case is to demonstrate that simply adding a strategic actor does not necessarily portend additional challenges for interpretation or identification. One could model the officer's behavior in different ways, for example by introducing some capacity constraint on investigation effort or changing the information structure of the game. This may change the interpretation of relevant reduced-form causal effects. Holding constant the sequence and selection into crime, however, changing the utilities or information of the officer cannot solve the identification problems described here.

4 When is a Theory Necessary for Identification?

The experiment and models in Section 3 provide some insights into how models of post-treatment interactions matter for identification and interpretation in the context of reporting and recording of crime. To what extent are these findings general? When are models of how a treatment impacts behavior necessary for identification?

4.1 Models of Post-Treatment Selection

A feature of the models in Section 3 is that strategies are chosen sequentially, not simultaneously: the crime occurs (resp. does not occur), then the bystander reports or does not report it, then it is investigated (resp. not investigated). Given the emphasis on sequence, I restrict attention to dynamic models.

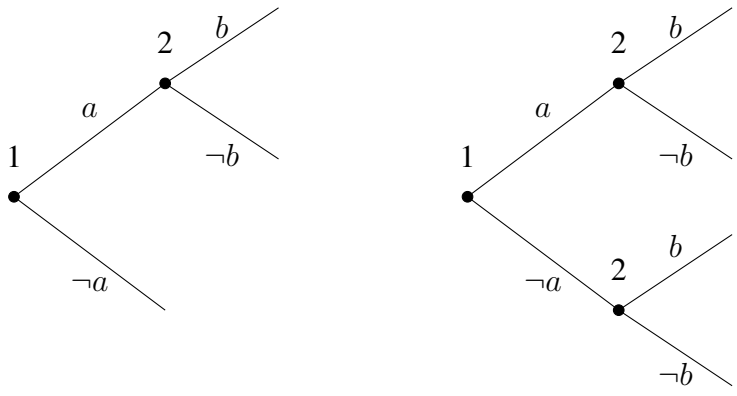
In describing dynamic models, I use the word "history" to mean the set of all previous post-treatment actions. As is standard, the set of histories (nodes) is denoted H. The first (post-treatment) node is $H^{\emptyset}$ and H^T represents a terminal node. In a static model, $H^{\emptyset} = H^T$. Adopting this notation, 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, a, is the same and has an equivalent strategy set, S_a, regardless of the strategy selected at h, for all $h \in H \backslash H^T$.*

Strategy set symmetry is straightforward to visualize in a game tree. Figure 3 depicts two games. On the left, Player 2's set of strategies, S_2, depends on the Player 1's action at the first node. As such, the strategy sets are asymmetric per Definition 1. In contrast, in the game on the right, Player 2's set of strategies, $S_2 = \{b, \neg b\}$, is equivalent regardless of Player 1's strategy at $H^{\emptyset}$.

Consider the connection between the game trees in Figure 3 and the DAGs in Figure 1. The asymmetric strategy set game tree (left panel) of Figure 3 is represented by panel (b) Figure 1. In contrast, the symmetric strategy set game tree is represented by panel (a) of Figure 1. Suppose that an experiment seeks to compare the difference in the frequency with which a population of Player 1's chooses a under some treatment Z. In either panel, so long as the Player 1's decision is measurable, one could estimate $E[a|Z = 1] - E[a|Z = 0]$, or the ATE of the treatment Z on the choice of a. In either panel (game) both potential outcomes are defined for all units.

Now suppose the researcher wants to understand the difference in the frequency with which a population of Player 2's chooses b under some treatment Z. In the left panel, this presents a



ASYMMETRIC STRATEGY SETS

SYMMETRIC STRATEGY SETS

Figure 3: Extensive form representation of simple dynamic games with (right) and without (left) symmetry in strategy sets.

problem. Player 2 does not act if Player 1 chooses $\neg a$. With abuse of notation, the potential outcomes $b(Z)$ and $\neg b(Z)$ are undefined if Player 1 selects $\neg a$. As such, $E[b|Z = 1]$ and $E[b|Z = 0]$ are undefined, rendering the ATE of Z on $b(Z)$ undefined. These potential outcomes are defined for individuals with history $H = a$. However, any comparison that conditions on the realization of Player 1's choice of a conditions on a post-treatment outcome. The researcher could seek to point- or interval-estimate the SACE, but the ATE is not identified.

In contrast, in the right panel of Figure 3, the ATE on Player 2's decision is identifiable under standard experimental identifying assumptions. The potential outcomes $b(Z)$ and $\neg b(Z)$ are defined, regardless of Player 1's decision. Importantly, the experimental research design used to manipulate Z can be identical in either panel of Figure 3. It is ultimately our assumptions about whether we are in the left or the right panel that determines whether the ATE on behavioral outcome b is identified. This observation suggests that theory is necessary for the identification of some estimands.

This paper proceeds to ascertain the conditions under which specification of such a theory is necessary. The findings on the minimal model in Figure 3 generalize to far more complex models of post-treatment behavior. The critical distinction for the identification of standard causal estimands, namely the ATE, depends largely on whether the theoretical model is strategy set symmetric. If

the model is not strategy set symmetric, the sequencing of the "selection" is of central importance for identification. In the framework developed here, "selection" occurs at any node, h, for which the actor or strategy sets at the following node depend on the action taken at node h.[10] Proposition 1 provides a general statement of this finding.

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. (Proof in Appendix.)

Proposition 1 provides several insights. Perhaps the most novel implication of Proposition 1 is that 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. The emphasis on causal identification 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 thus to find this variation in the assignment of some treatment; 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 this approach may not be consistent with the motivation of causal identification.

The primary threat to identification of the ATE identified by Proposition 1 is indeed post-treatment selection. Where this selection occurs in a sequence of post-treatment outcomes is critical. The ATEs of treatment on outcomes prior to and including the first instance of "selection" in

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

a sequence are identified. Subsequent to selection, the ATE is no longer identified. This finding posits a need for the specification of theory, particularly with respect to the analysis of so-called downstream outcomes of a treatment.

ATEs that are identified under Proposition 1 may or may not be substantively important for researchers. In some cases, this selection may simply measure treatment uptake. For example, consider a treatment that encourages citizens to initiate a bureaucratic process, e.g., registering for an ID or applying for a social program. Making the initial request may measure only "compliance" with treatment assignment, as opposed to a behavioral outcome of interest. Yet, under such a model, subsequent measures of participant interactions with the state are undefined for subjects that did not "opt in" in the first post-treatment action. Compliance with treatment assignment may or may not be of substantive import to researchers. As such, Proposition 1 does not guarantee the substantive importance of the identified estimands in a given context.

The invocation of a theory implies an increase in the amount, and possibly strength, of assumptions needed for causal identification. This imposition of stronger assumptions for identification is seemingly anathema to the research designs and estimators advocated by the identification revolution. In this context, thus, it is worth considering the implications of *not* specifying a theory. Following Proposition 1, claims of identification of ATEs on multiple behavioral outcomes in the absence of a theory imply several characteristics of an unspecified "shadow theory." Proposition 2 makes clear that authors must not be describing a dynamic model or the model must be strategy set symmetric with respect to identified outcomes. This implication of Proposition 2 suggests that researchers are not aided in this regard by theoretical agnosticism, even if the theory put forward is wrong.

Proposition 2. *In an experiment for which researchers claim to identify the ATE of $n > 1$ 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.*

Proposition 2 makes clear that in settings with multiple behavioral outcomes, claims of identification of an ATE (or ITT) cannot claim agnosticism as to theory. Given this finding, can specifica-

tion of an explicit theory of behavior actually *reduce* our concerns about the number or plausibility of the assumptions we invoke? We often seek to probe the empirical identifying assumptions through balance tests, placebo tests, or examination of parallel trends etc. To probe theoretical identifying assumptions, a clear statement of what assumptions are invoked for identification is necessary for assessment of the plausibility of these assumptions. In this sense, leaving such assumptions implicit *increases* our reliance on assumptions.

Further, suppose that researchers do not observe all behavioral outcomes of a treatment. This may because some outcomes are latent (like crime in the policing model) or because a researcher simply fails to measure an outcome of interest for any reason. In these cases, even if a researcher were to measure a single behavioral outcome, the intuition behind Proposition 2 may still apply. Specifically, if the researcher does not measure the first behavioral outcome, identification of the ATE on the measured outcome assumes the model is not dynamic or that all previous outcomes are measures of actions at strategy set symmetric histories. This danger is exemplified by the policing model in which the ATE on the first observed outcome – the bystander's reporting behavior – is undefined under the assumption of selection into crime.

4.2 When Should a Theory be Specified?

Proposition 1 implies that if a dynamic model is strategy set symmetric, then the ATE is identified for all post-treatment outcomes (under standard identifying assumptions). When, then, do we need to specify a theory for identification? One plausible approach would be to assume strategy set symmetry as a "null" or baseline state and justify deviations from such a model. Yet, there is no reason to believe that an assumption of asymmetry is rarer or less plausible than an an assumption of symmetry. To this end, I argue that as a baseline, there should always be an explicit account of post-treatment behavior when outcomes are sequential.

To this point, I have focused on dynamic decision- and game-theoretic dynamic models of complete information. To what extent does the argument generalize to other models? I consider static models and dynamic models with incomplete information.

Static models: First, consider a static model in which each player acts simultaneously. By def-

inition, a static game must be strategy set symmetric, since there is only one history ($H^{\emptyset} = H^{T}$). In the empirical setting of a static game, the dependent variable measures the strategy selected by each player(s) or some measure of the equilibrium outcome. Importantly, by definition, each player's actions are not contingent on any post-treatment history. In these settings, it is possible to identify the ATE on dependent variables measuring various aspects of player actions and "general equilibrium" outcomes. Nevertheless, a fully-specified theory is generally useful for interpretation of such empirical findings. In particular, when the dependent variable is some measure of equilibrium outcomes, the specification of a theory allows for a clear statement of expectations.

Incomplete information: Do dynamic models of incomplete information function differently than dynamic models of complete information? To answer this question, consider two empirical measures relevant to theories of this form: actions and beliefs. The implications for identification of outcomes measuring actions remains constant regardless of the information structure of the game. If a model is not strategy set symmetric, there must exist some form of post-treatment selection in the availability of strategies. The identification results in Proposition 1 persist in this case for the study of actions.

What do these results imply for the measurement and identification of outcomes measuring actors' *beliefs*? In general, at different nodes in a game of incomplete information, some beliefs are ruled out either in equilibrium or through full revelation of information. The identification question thus, is whether outcomes measuring beliefs that do not accord with theoretical predictions/assumptions are undefined. Based on the conception of unidentified outcomes in which unobserved outcomes exist on a dimension that is distinct from the measure of observed outcomes, this particular identification concern is absent for the study of measures of beliefs. If however, selection changes the composition of subjects that could feasibly have beliefs (i.e., through death or in the longer term through differential birth rates), identification challenges re-emerge. While these scenarios are present in some empirical settings, such compositional changes in the set of actors are not a standard feature of games of incomplete information.

A natural extension of consideration of beliefs includes other types of attitudinal outcomes,

i.e. elicited preferences. As in the case of beliefs, so long as the menu of options (e.g. the list of possible responses) for an attitudinal outcome does not depend on the post-treatment history, attitudinal outcomes do not introduce the same threat of post-treatment selection as sequential behavioral outcomes. Again, if the sample of subjects that could have preferences is a function of some form of post-treatment selection, familiar identification concerns return.

5 Implications for Research Design

The discussion to this point focuses on concerns on the challenges of the identification and interpretation of reduced-form experimental results in settings with sequential post-treatment behavior. Here, I turn to discussion of how these considerations should inform empirical research designs, proposing three classes of approaches in Table 3. These approaches have been used to varying degrees in existing applications, as shown by the citations in the table. However, these approaches have not been organized as a response to a common identification problem induced 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). I emphasize that these design recommendations are not mutually exclusive.

Panel A of Table 3 considers a treatment and outcomes as given and emphasizes the relevant estimands prior and subsequent to the first strategy set asymmetric history. This follows closely from the policing example. It is not yet standard practice to estimate the SACE. While Knox, Lowe, and Mummolo (2020) provide a new estimator and novel application related to policing, there are several limits what can be learned from a SACE. From a practical perspective, the SACE is often reported as an interval estimate produced by bounding estimators (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 equilibrium" effect (removing selection) Joffe (2011).

	Recommendation	Description in reference to asymmetric strategy set in Figure 3	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 3: Design recommendations. These recommendations refer to the left panel in Figure 3 (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.

Existing applications of the SACE typically occur outside strategic settings, often emphasizing squential 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 3, player 2's determination of b or $\neg b$ does not depend on player 1's decision. This follows directly from the logic of backward induction. As such, in strategic contexts, interval estimates of the SACE do not allow researchers to isolate the effect of treatment of player 2's actions.

Panel B of Table 3 suggests using multilevel random assignment to study sequential interactions. An ancillary experiment at the first history with an asymmetric strategy set could identify the ATE of a related but distinct treatment on Player 2's action, the determination of b or $\neg b$ given that Player 1 has played a. This approach is advocated by Green and Tusicisny (2012) in the context of lab experiments, and is exemplified by sequential multilevel experiments like Golden, Gulzar, and Sonnet (2019). This allows for identification of ATEs subsequent to some post-treatment selection. There exist two limitations to this recommendation. First and most practically, this strategy is likely limited to experimental (as opposed to quasi-experimental) settings and is infeasible in the context of some dynamic applications. Second, while 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 insight into the (general) equilibrium of by a model. Using this approach, the theoretical model indicates when a new manipulation is necessary for identification of the ATE.

Finally, Panel C suggests three changes in the measurement of outcome variables may help to address "truncation by death"-based selection issues. These strategies are generally simpler or cheaper to implement than ancillary experimentation. 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 $t + 1$. Whether this redefined outcome is relevant to the question or theory at hand will depend. Note that this suggestion departs from call to estimate incumbency advantage unconditional on running for office as advocated by De Magalhaes (2017). Indeed, the incumbency advantage unconditional on running imputes a "0" (or loss) for the undefined potential outcome (vote choice for a candidate that does not run). As in the discussion of truncation by death, this imputation implies a loss of information. The distinction between these approaches speaks to the importance of specifying the underlying interaction when redefining outcomes.

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.[11]

[11]Of course, latent crime incidence could be measured via crime victimization surveys or on-the-ground audits. However, the mapping between these measures and the administrative data may be non-trivial.

"Flattening" may be most attractive in cases like Findley, Nielson, and Sharman (2014) with a single actor (the agent). In a strategic settings, the flattened outcome may be determined jointly by multiple players' actions. In some case, outcomes measured in this way may measure equilibrium selection. Here, the focus is generally not the behavior of any single actor, but manifestations of some interaction. Specification of an equilibrium (or equilibria) is therefore quite important to the interpretation of flattened outcomes in strategic settings.

While changes in the measurement or definition of outcomes can permit identfication of the ATE in settings with post-treatment selection akin to truncation by death, I argue that these estimands cannot be interpreted without a minimal theory of the post-treatment causal process. When researchers employ these design choices, they employ them as a function of an underlying theory. An extensive form shows researchers where to "stop" in terms of identification of the ATE on behavioral outcomes. It also guides researchers in determining where to flatten or redefine outcomes. Some of the resultant measures are less obvious outcome measures than individual actions. Here, theory serves to provide justification for identification of the ATE and helps to rationalize the specific outcomes.

5.1 Generalization from Experimental to Observational Designs for Causal Inference

To this point, I have focused on experiments and identification of the ATE or the ITT. Yet, the argument applies more broadly to other research designs and estimands. Similar arguments are relevant to the local ATEs (LATEs) or average treatment effect on the treated (ATT) estimands emphasized in popular observational designs for causal inference, as suggested by the examples in Table 1. Two features of observational studies heighten concerns about post-treatment selection in observational settings: less direct observation of the post-treatment causal process and longer post-treatment histories.

First, consider the degree of researcher observation of the causal process. Recall that Proposition 1 holds that, under standard identification assumptions, the ATE of treatment on an actor's behavior at $H^{\emptyset}$ is identified. When researchers can observe and measure all behavioral outcomes (as in the lab), they can estimate at least one ATE (or, by extension, LATE or ATT). However,

when observation of this process is less comprehensive, it is possible that among observed behavioral outcomes, the ATE is never identified, as in Cases #2-#4 of the policing example. This would occur if all measured outcomes are realized after the first strategy set asymmetric history. Researchers leveraging observational designs may be less able to systematically observe a sequence of behavioral outcomes than researchers employing experimental designs.

For example, in an experimental intervention in which researchers design implementation of treatment and data collection, there may be more room for observation – qualitatively or quantitatively – of how various actors respond to a treatment. For example, some experiments on electoral accountability in the recent Metaketa-I find evidence of a measurable response by political campaigns (Dunning et al., 2019). It is less clear that authors would have the ability to detect or measure these responses in an analogous observational study (e.g., Ferraz and Finan, 2008).

Second, when observational work invokes a longer causal chain, a theory of downstream outcomes is apt to be more "involved" than a theory explaining an initial behavioral outcome. Given the reliance on strategy set symmetry to identify ATEs of sequential outcomes, a longer sequence of behaviors with claims to identification requires that an extensive form maintain this structure over more histories. This can be considered a stronger (or more restrictive) theoretical assumption.

Combining these observations, in settings researchers have less ability to observe what happened during and after the implementation of the "treatment," the sequencing of interactions can be less self-evident. In considering (1) the strength of theoretical assumptions needed for identification; and (2) our ability to observe the underlying process, it may be the case that the settings that most need theory to ground identification are precisely those in which we must rely upon the strongest and least testable assumptions.

6 Conclusion

This paper considers challenges to causal identification that emerge in studies with multiple, potentially sequential behavioral outcomes. I show that standard estimands are identified by a research design for specific outcomes. The finding that identification is relative to an outcome suggests

a need to impose some assumptions (structure) about the post-treatment causal process if causal inference is a goal. Toward this end, applied theory is necessary to ground claims of identification in empirical settings with sequential outcomes.

A natural objection to this position asks what happens if a theory is wrong. To this I provide two responses. First, most theories are "wrong" in some respect. However, the minimal notion of a theory implied by Proposition 1 is simply a sequence of actors and strategies, absent utilities, or even an equilibrium concept. In some contexts, particularly in studies with short histories, this sequence is observable to researchers, which can help to ground assumptions using qualitative or quantitative evidence. Certainly, interpretation concerns hinge on how a researcher models preferences, the type of model, and, where relevant, the equilibrium concept. The good news is that the identification concerns here are somewhat less exacting in terms of model specification than those of interpretation.

Second, following Proposition 2, the absence of a theory makes (possibly strong) assumptions about the sequence of actors and strategies. Namely, it assumes that the extensive form of a game is strategy set symmetric, at least through the measured outcomes. If this is the case, enumerating the implicit theory allows researchers to shed light on their assumptions and provides grounds for probing such assumptions more explicitly. In other words, even if one views agnosticism as a virtue in the context of empirical research, the absence of a theory in the context of multiple behavioral outcomes should not be equated with theoretical agnosticism.

The ultimate insights of this paper provide guidance for empirical research design. 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 of causal inference. Ultimately, this paper calls for a more explicit marriage of theory and data in identification-oriented empirical work.

References

Abramson, Scott F., Korhan Koçak, and Asya Magazinnik. 2021. "What Do We Learn about Voter Preferences from Conjoint Experiments?" Working paper available at https://www.dropbox.com/s/kracengw1mvz5my/Conjoint_AJPS_Submission%20%281%29.pdf?dl=0.

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.

Arias, Eric, Rebecca Hanson, Dorothy Kronick, and Tara Slough. 2019. "The Construction of Trust in the State: Evidence from Police-Community Relations in Colombia." Pre-Analysis Plan.

Aronow, Peter M., and Benjamin T. Miller. 2019. *Foundations of Agnostic Statistics*. New York, NY: Cambridge University Press.

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.

Dunning, Thad, Guy Grossman, Macartan Humphreys, Susan D. Hyde, Craig McIntosh, and Gareth Nellis, eds. 2019. *Information, Accountability, and Cumulative Learning: Lessons from Metaketa I*. New York: Cambridge University Press.

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.

Ferraz, Claudio, and Frederico Finan. 2008. "Exposing Corrupt Politicians: The Effects of Brazil's Publicly Released Audits on Electoral Outcomes." *Quarterly Journal of Economics* 123 (2): 703–745.

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.

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, 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. 2013. "Is Theory Getting Lost in the "Identification Revolution"?" *The Political Economist: Newsletter of the Section on Political Economy, American Political Science Association* X (1): 1:3.

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

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.

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.

Manski, Charles E. 1995. *Identification Problems in the Social Sciences*. Cambridge, MA: Harvard University Press.

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.

Morton, Rebecca B. 1999. *Methods and Models: A Guide to the Empirical Analysis of Formal Models*. New York: Cambridge University Press.

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.

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`.

Sun, Jessica S., and Scott A. Tyson. 2019. "Theoretical Implications of Empirical Models: An Application to Conflict Studies." Working paper.

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.