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

Tara Slough*

October 30, 2020

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 post-treatment 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 multiple 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, 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 (Clark and Golder, 2015; Ashworth, Berry, and Bueno de Mesquita, 2015; Samii, 2016; Huber, 2017; Franzese, 2020). 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 multiple behavioral outcomes, I characterize 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 identificaton.

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 Pishke, 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. Yet, the structure of what happens *after* treatment also poses underappreciated limits to identification of causal estimands, beyond typical discussions of non-interference (SUTVA). I argue that theory provides necessary assumptions to structure thinking about responses to treatment. In a setting with (possibly) strategic actors, these considerations are particularly important. To that end, this paper asks: under what conditions must researchers impose additional assumptions from applied theory in order to identify causal estimands even with "credible" research designs? What are the costs of misspecifying the theory for our ability to identify and interpret causal effects?

Consider two general approaches to estimating causal effects with differing roles for theory. First, an "econometric" or structural approach to causality involves an explicit modeling of agents' preferences and behaviors that generate observable data (Heckman, 2008). This involves an ex-

plicit modeling of how treatments are allocated (what happens "before" treatment) and how actors respond (what happens "after" treatment). With these models and data, researchers can study the causal process and its implications by estimating structural parameters if an estimator can be derived. With this approach, theory dictates the empirical strategy. In so doing, the theory introduces a (possibly large) set of assumptions necessary to estimate causal parameters.

In the second approach, typified by research designs invoking the Neyman-Rubin causal model, theory and empirics are not necessarily intertwined. This view of causality dominates current empirical research in political science. The assumptions underlying the identifcation of causal estimands focus strongly on how treatment was assigned and delivered to subjects. To the extent that these designs are implemented to test the implications of theories, causal estimands offer reduced-form tests. These estimands (e.g., the average treatment effect [ATE]) rarely, if ever, represent parameters of theoretical models. Nor does identification of these estimands permit the identification of underlying theoretical parameters.

In this paper, I argue that theory is necessary for identification and interpretation in many research designs in the Neyman-Rubin tradition. One natural response this position is simply to "go structural," or adopt the former view of causality. Structural models in political science remain relatively rare and disciplinary distinctions between economics and political science limit, to some extent, the application of structural models to the study of politics.[1] Political science generally lacks underlying organizational principles with direct (numerical) analogues in the data like the supply and demand framework or models of choice in economics. To the extent that political science exists in latent concepts (i.e., the ideological spectrum), the mapping from theoretical parameters to data poses an additional hurdle. Given these limitations in some applications and the manifest popularity of design-based approaches, investigating the role of theory in design-based research offers practicable insights.

I focus on the role of theory in research designs that study a sequence of outcomes. Given the logistical difficulties of manipulating a treatment or finding an "as if" randomly assigned treat-

[1]For examples of structural work in political science, see Kalandrakis and Spirling (2011), Crisman-Cox and Gibilisco (2018), and Abramson and Montero (2020).

ment in the world, we often seek to study the effects of a treatment on multiple, often sequential, outcomes. This paper contends that the relationship between reduced-form causal estimands and theoretical predictions can be particularly ambiguous in settings with sequential outcomes. Theories that organize assumptions and generate predictions about "what happens after treatment" provide interpretation necessary to interpret results in light of such ambiguity. I show that the invocation of different theoretical assumptions with the same research design imply the identification of different estimands in settings with sequential post-treatment outcomes. This analysis suggests that there exist research designs in which it is necessary to specify theory for valid causal identification even when standard identifying assumptions hold.[2]

I define a theory as a model, or an abstract representation of the world, that relies upon deductive reasoning (Clarke and Primo, 2012). The type of model is not crucial to the argument advanced in this paper. In practice, models could be decision theoretic, game theoretic, behavioral, social choice, or computational, among others. The crucial requirement of a theory in this context is that it models: (1) relationships between an exogenous treatment and endogenous outcomes; and (2) relevant relationships between endogenous outcomes. A basic reading of the potential outcomes framework may suffice for the former, but does not typically incorporate dependencies between outcomes.[3]

This paper makes three contributions. First, it clarifies the ubiquity of post-treatment selection in identification-oriented research designs and its consequences for identification of standard causal quantities. Second, it derives conditions on the structure of post-treatment behavior under which the ATE and, by extension, other standard estimands are theoretically identified. Finally, it shows that an absence of an explicit theory does not necessarily imply greater "agnosticism" about the causal process. These contributions suggest the need for more explicit mapping between applied

[2]Throughout this paper, I refer to "standard identifying assumptions" as the minimal set of assumptions invoked for identification of a causal estimand given a research design. For example, the standard identifying assumptions for the intent to treat effect in an experiment are: ignorability of treatment assignment; excludability of treament; and SUTVA (Green and Gerber, 2012).

[3]As I show in Section 3, the implications of more complex models can be usefully be mapped back into the potential outcomes framework.

theory and claims to identification of reduced-form estimands in identification-oriented research.

My argument builds upon 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 a published empirical finding. These models guide discussion of the conditions under which reported estimates provide evidence supportive of the theoretical claims advanced (e.g., Ashworth and de Mesquita, 2014; Izzo, Dewan, and Wolton, 2020; Prato and Wolton, 2019; 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 – multiple 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.

The implications of this paper speak broadly to literature on identification-driven research designs. Most specifically, 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 range of applications.

Finally, this paper develops an ongoing debate about the relationship between theory and causal identification. I complement discussion of the compatibility between specific empirical settings and goals of causal identification in certain substantive domains (e.g., Binder, 2019; Luna, Murillo, and Shrank, 2014) by proposing a class of theoretical settings (dynamic models) in which these concerns are critical. I further argue that identification-driven research designs should not and, in some settings, cannot be separated from theory, building on broader debates about the relationship between theory and causal identification (e.g., Huber, 2013; Samii, 2016).

2 Definition, Identification, and Interpretation of Causal Estimands

2.1 Multiple Outcomes

Researchers often study the effects of some treatment (or independent variable), Z, on more than one outcome. Multiple outcomes might be used to test multiple observable implications of a theory, test competing arguments, or to address welfare effects in a program evaluation context. The identification revolution has sharpened the focus on the data generating process governing the assignment of Z. In contrast, comparatively less attention is devoted to the relationship between outcome variables.

Figure 1 depicts three possible relationships between a randomly assigned treatment, Z, and two outcomes, Y_1 and Y_2.[4] In Panel (a), while both outcome variables have a common parent, Z, neither outcome is a function of the other. In potential outcomes notation, we could write $Y_1(Z)$ and $Y_2(Z)$. In contrast, in Panel (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 $Y_1(Z)$.

The difference in functional relations between variables in (a) and (b) may lead researchers to utilize other estimators, including those suggested by mediation analysis in Panel (b). In this paper, the focus is instead about the difference between panels (b) and (c). Like Panel (b), Panel (c) of Figure 1 suggests that Y_2 is a function of both Z and Y_1, but 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 not defined. Such undefined outcomes undermine claims of causal identification (Holland, 1986). I argue that extensions of panel (c) 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

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

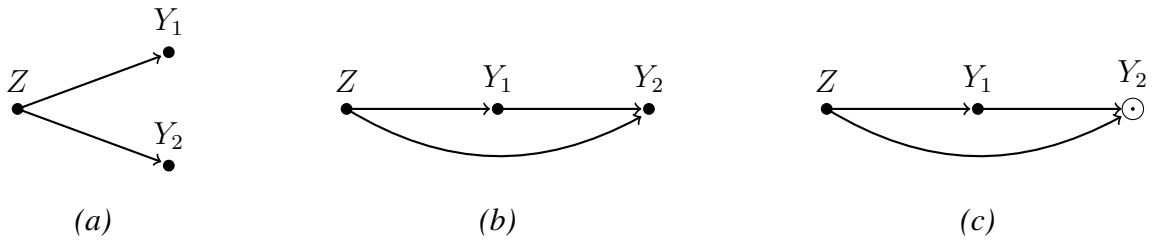


Figure 1: Three graphical causal models of the relationship between a randomly assigned treatment, Z, and two outcomes, Y_1 and Y_2. The $\odot$ node indicates that outcome variable Y_2 is not defined for all levels of outcome variable Y_1.

causal inferences can be separated into identification and statistical components (p. 4). 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 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 six literatures that often make claims to causal identification: long-run development, conflict, email audits, and incumbency advantage using experiments, natural experiments,

[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	Treatment	Outcome	Post-treatment selection
1	Long-run development	Imposition of colonial institutions in colonial-era communities	Individual or community-level development outcomes in present communities	Community non-persistence from colonial era to present or differential rates of survival, out-migration.
2	Effects of conflict	Community exposure to conflict	Individuals' political attitudes or behaviors	Death during conflict.
3	Email audit experiments	Petitioner/petition characteristics	Quality of response (accuracy, respect etc.)	Subject does not respond to email.
4	Ideological positioning	Electoral performance, t	Platform (ideology) in election $t+1$	Party ceases to exist in election $t+1$
5	Incumbency (dis)advantage	Incumbency	Vote share of incumbent candidate or party in election $t+1$	Candidate does not run in election $t+1$.
6	Police use of force	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. Some of these issues are discussed in existing literature including: incumbency advantage (e.g., Erikson, 1971; Erikson and Titiunik, 2015; Eggers, 2017), audit studies (Coppock, 2019; Slough, 2020), and policing (Knox, Lowe, and Mummolo, 2020).

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 Sequencing of Outcomes and Interpretation of Reduced-Form Estimates

Reduced-form estimates of causal estimands give rise to myriad questions of interpretation. Research designs capable of identifying a causal estimand often lack the ability to ascertain *why* we observe an effect. Consider an exogenous treatment representing a shock to the value of a single theoretical parameter, which in turn drives (possible) differences in an actor's behavior. Estimates of the causal effect of the treatment on this behavior, in general, do not permit identification of the underlying parameter. As such, reduced-form tests of a theory rely upon estimation of quantities that are related to, but ultimately epiphenomenal to the theory.

The relationship between causal estimands and underlying parameters can be particularly ambiguous in the context of theories with sequential outcomes. Sequential outcomes imply some form of dynamic model. Whether a single actor makes a sequence of decisions or multiple players interact in sequence, the implications of a change in an exogenous parameter on causal estimands of interest is not necessarily clear without a set of assumptions about the causal process of interest. In particular, if anticipation of future actions drives players' actions, a treatment that manipulates one theoretical parameter can affect observed outcomes through multiple channels. This can lead to ambiguity about the predicted sign of causal estimands or ambiguity as to which channels are at work. For the purposes of interpretation, specifying theoretical assumptions and predictions allows for a clear statment about what implications of a theory a causal estimand could be testing.

2.5 Alternative Estimand and Relation to 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 $S(Z)$ represents the post-treatment selection outcome, 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 fuller exposition of this principal stratification approach, see Appendix A. 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 (here, an equilibrium) 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 and illustration 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 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. A difference-in-means estimator 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.

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

[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.

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.

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.[9] 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

[9] Alternatively, it counters social norms against reporting. For this reason, I consider this cost as the net cost of reporting relative to not reporting.

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: $0 < \alpha_{\neg r} < \alpha_r < s$.

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 B.

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 $v \cap i$.

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.

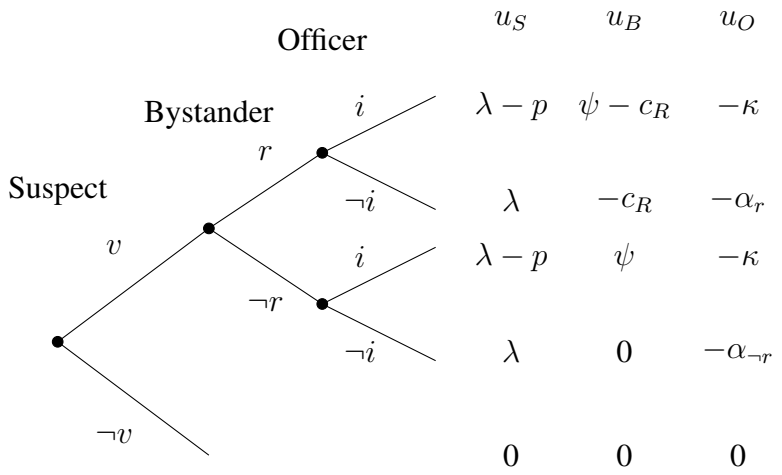


Figure 2: Extensive form representation for Case #4.

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 B, 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 $SACE$s 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. 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 countervailing: 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 not possible to sign the resultant difference-in-means estimates.

As in Case #2, selection into crime renders both ATEs undefined. The $SACE$s 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 $\tilde{\lambda}$ where the suspect is indifferent between committing the crime and not committing the crime when $Z = 1$. 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, $\underset{\sim}{\lambda} \leq \tilde{\lambda}$. 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 this 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 B. 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.[10] 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

[10]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.

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, $\iota_R > \iota_N$ approximate the officer's equilibrium strategy.

As in both cases where there exists some form of selection into crime, the relevant ATEs are undefined. The $SACE$s 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. I relegate the formal statement of these results along with the proofs to Appendix B.

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.

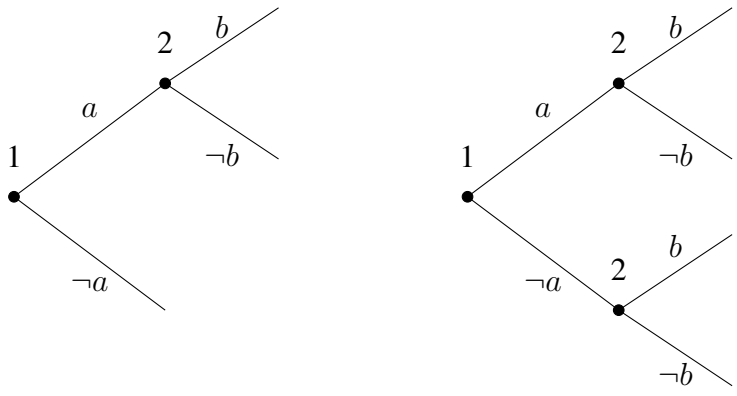
In describing the the importance of a dynamic models, I use the word "history" to mean the set of all previous post-treatment actions. As is standard, the set of histories (nodes) is denoted H. The first (post-treatment) node is $H^{\emptyset}$ and H^T represents a terminal node. In a static model, $H^{\emptyset} = H^T$. 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 is the same and has an equivalent strategy set 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 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, $\{b, \neg b\}$ are 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 (c) Figure 1. In contrast, the symmetric strategy set game tree is represented by panel (b) 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-identify 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.[11] 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 and, 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 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

[11] 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. In addition to the standard empirical assumptions justifying the research design, we add assumptions about how actors behave (and why) that are needed to justify empirical causal claims. In this sense, many existing claims of causal identification in the applied social science literature rely on the validity of an underlying (implicit) model of behavior. The argument here is simply that by making this model and related assumptions explicit, it is possible to determine which estimands are identifiable in a given design.

The 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 specification 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 necesssary for assessment of the plausibility of these assumptions. In this sense, leaving such assumptions implicit *increases* our reliance on assumptions.

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.

Beyond the identification concerns outlined here, the invocation of a theory is useful for the interpretation of causal estimands. Particularly in the case of multiple behavioral outcomes, theory can help to decompose (if not identify) causal channels via which a treatment should affect the reduced-form estimands we seek to estimate. For example, even in the simplest decision theoretic model in Section 3, the theory provided a clear, if counterintuitive prediction that we should observe more recorded instances of crime in the administrative data if crime is not changed by treatment. To that end, theory can be helpful beyond justifying claims to identification of causal estimands.

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 definition, 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),

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

5.1 Best Practices for Experimental Design

The discussion to this point focuses on concerns on the challenges of the identification and interpretation of reduced-form experimental results. Here, I turn to discussion of how these considerations should inform experimental design, proposing two approaches. Identification of the ATE (or ITT) is possible under a theoretical model; the absence of such a model does not make an analysis strategy of sequential outcomes any more "agnostic." Instead, it implies strong assumptions about the extensive form of an (unspecified) model. Even if the assumption of strategy set symmetry seems plausible in the absence of an extensive form, the absence of a model provides little guidance for interpretation of the resultant $ATEs$.

Given the left panel of Figure 3 depicting the non-strategy set symmetric game, one could imagine two approaches to experimental design and analysis. The first approach holds constant the design. In this context, researcher can identify the ATE on the first action (the determination of a or $\neg a$). That decision may well incorporate consideration of Player 2's actions, as we typically would assume in characterizing an SPNE. In so doing, the model may provide additional implications for testing, even in the absence of estimating any causal effect on Player 2's action. The model provides two benefits in this setting. First, the model (or assumptions therein) defines

the set of defined estimands. This informs researchers' choice of outcomes. Second, it may imply provide additional tests of the theory's implications.

A second approach would be to add a second or ancillary experiment to identify the ATE of a related but distinct treatment on Player 2's action, the determination of b or $\neg b$ when 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. It permits 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 model implies when a new manipulation is necessary for identification of the ATE.

5.2 Redefinition and Selection of Outcomes

To this point, I fixed the behavioral outcomes of interest and use a model to suggest the set of identified estimands. One alternative is to redefine the outcomes of interest in order to eliminate the identification issues stemming from undefined potential outcomes. Depending on the design, three possibilities include: (1) measurement of more outcomes prior to the variable measuring the first strategy set asymmetric history; (2) redefinition of potential outcomes to reduce the threat of selection; and (3) "flattening" of a sequence of actions into a categorical outcome. Note, however, that characterization of an extensive form is critical to the determination and viability of these strategies. Moreover, each strategy involves redefining (or adding to) the set of identified estimands; whether these quantities answer the research questions of interest remains an open question.

First, in the context of "truncation by death," researchers often search for clinical markers that present quickly, ideally prior to death (selection). Thus, the set of outcome(s) of interest thus can then be augmented to include outcomes unlikely to be undefined as a function of other post-treatment outcomes. In the social science setting, we can gain leverage by measuring outcomes that present prior to the first non-strategy set symmetric history. For example, if a threat to the validity of a study of incumbency advantage is that the challenger will not contest election $t + 1$,

a public opinion poll about a hypothetical race between the incumbent and the (same) challenger prior to candidacy announcements provides potentially relevant evidence that does not suffer from post-treatment selection like vote choice. Importantly, these outcomes are distinct from the original behavioral outcomes of interest. Whether such outcomes are sufficient to answer a question or test a theory depends on the context.

Second, researchers may redefine the potential outcomes of interest to limit 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.

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. Such "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. To interpret flattened outcomes as a measure of equilibrium selection, one needs to specify the equilibria to understand the mapping between these flattened outcome measures and theoretical predictions. This type of analysis will require a complete model as opposed to simply an extensive form.

5.3 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. Informal discussions of the "credibility" of methods for drawing causal inferences tend to focus on the plausibility of identifying assumptions. For example, we often consider researcher control over the assignment of treatment as one feature that makes claims of ignorability of treatment assignment more plausible. This article studies variability in the identification of estimands even when standard identifying assumptions hold. The natural analogue of these discussions in the context of post-treatment selection thus considers the limits of our ability to observe or model post-treatment behavior.

What provides researchers leverage to accurately model and study these post-treatment behaviors? I suggest two dimensions upon which research designs vary with implications for researcher knowledge about the post-treatment causal process. First, research designs vary in the possible length of the post-treatment "history." Consider two extreme research designs intended to estimate causal effects. On the short extreme, survey experiments vary the prime that subjects read immediately before reporting a belief, attitude, or hypothetical behavior. The design effectively ensures that we measure an outcome with history $H^{\emptyset}$. On the other extreme, deep history or institutional origins "natural experiments" have long histories by definition. To the extent that past actions condition the set of available strategies, researchers should be particularly skeptical about how these causal processes limit our ability to identify causal effects on downstream outcomes.

The finding that a research design may identify an estimand for some outcome(s) but not others(s) suggests that identification of the ATE (or other causal estimand) on early outcomes in a post-treatment causal chain is less likely to be problematic than identification of later outcomes. For this reason, theoretical assumptions about the causal process may be particularly important for understanding the effects of a treatment on downstream outcomes.

Given a longer causal chain, a theory of downstream outcomes is apt to be more "involved" than a theory for an initial outcome. Such theories invoke more assumptions toward identification of an estimand. Our ability to validate such assumptions about the causal process, however, may

depend on our ability to observe the underlying causal process. Research designs vary substantially in researchers' ability to observe such post-treatment behavior. For example, in an experimental intervention in which researchers design implementation of treatment and data collection, there is often 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).

In contrast, in many natural experiments, researchers are confined to data and observations collected from other sources. While original (non-archival) outcome data collection is often helpful, specifying an extensive form can be more challenging. When we 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 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., and Sergio Montero. 2020. "Learning about Growth and Democracy." *American Political Science Review* Forthcoming.

Angrist, Joshua D., and Jorn-Steffen Pishke. 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.

Binder, Sarah. 2019. "How we (should?) study Congress and history." *Public Choice* pp. 1–14.

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.

Clarke, Kevin A., and David M. Primo. 2012. *A Model Discipline: Political Science and the Logic of Representations*. New York, NY: Oxford University Press.

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.

Crisman-Cox, Casey, and Michael Gibilisco. 2018. "Audience Costs and the Dynamics of War and Peace." *American Journal of Political Science* 62 (3): 566–580.

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.

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`.

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`.

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

Kalandrakis, Tasos, and Arthur Spirling. 2011. "Radical Moderation: Recapturing Power in Two-Party Parliamentary Systems." *American Journal of Political Science* 56 (2): 413–432.

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.

Luna, Juan Pablo, María Victoria Murillo, and Andrew Shrank. 2014. "Latin American Political Economy: Making Sense of a New Reality." *Latin American Politics and Society* 56 (1): 3–10.

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.

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

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.

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.

Appendices

A Formal Exposition of Truncation by Death Problem

Consider the following graphical model:

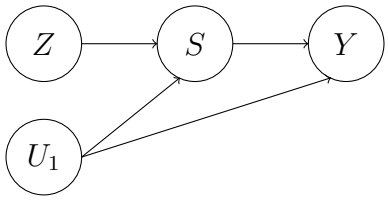


Figure 4: A graphical model depicting a causal process consistent with endogenous units of analysis if the potential outcome $Y(Z, S)$ is undefined for some units on the basis of the revelation of some $S(Z)$.

Treatment $Z_i \in \{0, 1\}$, is assigned such that the probability of assignment to treatment $Z = 1$ is $p \in (0, 1)$ for all units. A first outcome, $S(Z) \in \{0, 1\}$ indicates whether the a subject "survives." The dependent variable of interest $Y(S, Z)$ occurs subsequent to the realization of $S(Z)$. Define four causal types (principal strata): always survivors, if treated survivors, if untreated survivors, and never survivors. Table 3 defines these types, their shares in the population, and relevant potential outcomes.

Stratum	Weight	$S(Z=1)$	$S(Z=0)$	$\bar{Y}(S=1, Z=1)$	$\bar{Y}(S=1, Z=0)$	$\bar{Y}(S=0, Z=1)$	$\bar{Y}(S=0, Z=0)$
Always survivor	π_A	1	1	$\bar{Y}_A(1,1)$	$\bar{Y}_A(1,0)$	-	-
If **T**reated survivor	π_T	1	0	$\bar{Y}_T(1,1)$	-	-	$[\bar{Y}_T(0,0)]$
If **U**ntreated survivor	π_U	0	1	-	$\bar{Y}_U(1,0)$	$[\bar{Y}_U(0,1)]$	-
Never survivor	π_N	0	0	-	-	$[\bar{Y}_N(0,1)]$	$[\bar{Y}_N(0,0)]$

Table 3: Principal strata of an experiment with a binary treatment and binary survival variable. Elements in brackets indicate that a potential outcome is undefined. If defined, the outcome $Y(S, Z) \in \mathbb{R}^1$ and the last four columns indicate cell means.

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

$$E[Y(Z=1)] - E[Y(Z=0)] = \begin{array}{l} \pi_A\bar{Y}_A(1,1) + \pi_T\bar{Y}_T(1,1) + \pi_U\underline{\bar{Y}_U(0,1)} + \pi_N\underline{\bar{Y}_N(0,1)} - \\ (\pi_A\bar{Y}_A(1,0) + \pi_T\underline{\bar{Y}_T(0,0)} + \pi_U\bar{Y}_U(1,0) + \pi_N\underline{\bar{Y}_N(0,0)}) \end{array} \quad (2)$$

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

B Equilibrium Characterization, Proofs from Stylized Models

B.1 Case #1: Always Crime

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

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

Solving for ψ, the citizen will report if:

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

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

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

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

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

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

B.2 Case #2: Exogenous Crime and Exogenous Investigation

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

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

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

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

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

B.3 Case #3: Endogenous Crime and Exogenous Investigation

I characterize a subgame perfect equilibrium in pure strategies by backward induction. As such, I begin with the citizen's decision whether or not to report a crime in the subgame in which a crime has occurred. This is equivalent to the citizen's calculation in subsection B.1. The citizen reports if and only if:

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

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

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

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

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

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

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

treatment and control:

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

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

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

A difference-in-means estimator estimates:

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

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

B.4 Case #4: Strategic Policing

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

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

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

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

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

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

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

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

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

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

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

the remark follows directly from the proof to Remark 3.

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:

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

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

Both expressions are ambiguous in sign.

■

C Proof for Proposition 1

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

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

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

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

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

$$\sum_{h \in H^j} Pr(h|Z = 1) E[\mathcal{A}|Z = 1, h = h] - \sum_{h \in H^j} Pr(h|Z = 0) E[\mathcal{A}|Z = 0, h = h] \tag{3}$$

First, consider the first post-treatment action, $j = 0$. Both expectations in Equation 3 are defined. The ATE is both defined and identifed given Assumptions 1 and 2 and standard arguments (i.e. Green and Gerber (2012) Equation 2.3 or Angrist and Pishke (2010) Section 2.2).
Now, consider some $j > 0$. Consider two9 cases:

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

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