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

Tara Slough*

May 13, 2019

*Preliminary. Please download the latest version here.*

**Abstract**

What is the role for theory in identification-driven research designs? In this paper, I argue that not only is theory important for the interpretation of causal findings, but in many cases, theory is necessary for the identification of standard causal estimands (i.e., the ATE). In particular, I show that when empiricists study a sequence of post-treatment behavioral outcomes, post-treatment selection can prevent the identification of standard causal estimands, even when standard (empirical) identification assumptions hold. In these cases, articulation of a theory, or model of the world, that defines the post-treatment selection processes is necessary to define a set of identified causal estimands. Using a stylized example of crime, reporting, and recording, I illustrate how theory is necessary to reveal the set identified causal estimands, holding constant the research design. I then present this result more generally by considering the conditions under which the invocation of a theory is necessary for causal identification. I consider the implications for different research designs. This paper illustrates the need for theory in many identification-driven research designs.

*Ph.D. Candidate, Columbia University and Predoctoral Fellow, University of California, Berkeley, tls2145@columbia.edu. I thank Danny Hidalgo, Winston Lin, Fredrik Sävje, and audiences at the Harvard Experiments Working Group and Yale Quantitative Methods Seminar for helpful comments. This project is supported in part by an NSF Graduate Research Fellowship, DGE-11-44155.

# 1 Introduction

"Is theory getting lost in the 'identification revolution'?" This question has arisen repeatedly in political science since posed by Huber (2013). The "identification revolution" refers to the popularization of methods of design-based inference that are employed to identify causal effects. More recent reflection on the relationship between theory and identification-driven empirical resesearch in political science centers on whether or not there exist tensions between the goals of theory and identification (Clark and Golder, 2015; Ashworth, Berry, and de Mesquita, 2015; Samii, 2016; Huber, 2017). This paper reconsiders the role of applied theory in the context of design-based research in political science, positing theory as a necessity for identification of standard causal estimands in a large set of applications.

The "identification" or "credibility" revolution in empirical social science places an increasing emphasis on research design (Angrist and Pishke, 2010). In particular, researchers turn to research designs that invoke fewer and/or less heroic modeling assumptions to identify causal estimands (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 treatment. Yet, the structure of what happens *after* treatment also poses underappreciated limits to causal identification, 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 and interpret 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.[1]

[1] I omit discussion of graphical models at the moment (e.g., Pearl, 2009). The argument is not inconsistent invocation of graphical models for identification. Some components of the discussion may be clearer through such models rathern than the potential outcomes framework.

First, an "econometric approach" or structural approach to causality involves an explicit modeling of agents' preferences and behaviors that generate observable data (Heckman, 2008). This involves an explicit 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, given that 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 estimation of causal effects are generally empirical. To the extent that these designs are implemented to test the implications of theories, causal estimands provide reduced form tests. In general, these estimands (e.g. the average treatment effect) never represent parameters of theoretical models. Nor does estimation of these estimands permit the identification of underlying parameters in a theory.

This paper makes the case 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. While the number of structural models in political science has increased in recent years (see e.g., Kalandrakis and Spirling, 2011; Crisman-Cox and Gibilisco, 2018), disciplinary distinctions between economics and political science impose limitations in the application of structural models to the study of politics. Political science generally lacks underlying organizational principles with direct (numerical) analogues in the data. Structural models in economics are often built on supply and demand framework or some model of choice. 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 set of assumptions. Given these limitations, investigating the role of theory in design-based research offers practicable insights.

The core insight of this paper is that theory is often necessary to understand what happens

after the administration of "treatment." Standard identifying assumptions,[2] with the possible exception of the stable unit treatment value assumption (SUTVA), focus on what happens before the treatment (e.g. parallel trends in a difference-in-differences), how treatment is assigned (e.g. ignorability or random assignment), and the relationship between treatment assignment and the receipt of treatment (e.g. first stage, excludability, and monotonicity in an IV). Yet, even if these assumptions hold, there is no guarantee that an estimand is identified for every post-treatment outcome. Recent literature describes such problems of post-treatment selection or partial identification in the context of audit experiments (Coppock, 2019; Slough, 2018) and policing (Knox, Lowe, and Mummolo, 2019). This paper contends that in general, a further set of assumptions from applied theory are necessary to understand what estimands are defined, a precondition for identification, for which post-treatment outcomes.

I define a theory as a model, or an abstract representation of the world, that relies upon deductive reasoning (Waltz, 1979; Clarke and Primo, 2012). The type (technique) 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 examples in this paper are decision and game theoretic, but the framework admits other types of theories.

The argument forwarded in this paper relates closely to a recent literature considering the "theoretical implications of empirical models" (TIEM). The most common approach to TIEM involves writing a model to interpret a published empirical finding. Then, modelers use the model to understand whether estimates provide evidence for the claims forwarded (e.g., Ashworth and de Mesquita, 2014; Prato and Wolton, 2019; Izzo, Dewan, and Wolton, 2018). A second approach examines a research design and setting to examine the validity of the design on the basis of an underlying theory (e.g., Eggers, 2017). This paper provides an argument following the second approach, making generalizations to a wider array of research designs.

The implications of this paper speak broadly to literature on identification-driven research designs in the Neyman-Rubin causal framework. Most specifically, the focus on what happens after

[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.

treatment represents an increasing concern in research design. Yet existing works largely focus on the ills of "bad" controls (Montgomery, Nyhan, and Torres, 2018) or specific types of post-treatment selection (Knox, Lowe, and Mummolo, 2019). The treatment in this article is much more general and invokes theory as necessary to identify the possibility of post-treatment biases in some cases. The focus on theory as a necessary component of at least some research designs suggests limitations of current expositions of what constitutes a minimal research design (Blair et al., 2016).

Finally, this paper contributes to an ongoing debate about the relationship between theory and causal identification. In general, I find no tension between the use of identification-driven research designs and the invocation of a theory, contra (Huber, 2013, 2017). However, if researchers opt to use specific research designs *because* there is less need to state an argument explicitly, we should observe tensions of the sort identified by (Huber, 2013, 2017).

## 2 Definition, Identification, and Interpretation of Causal Estimands

### 2.1 Undefined Potential Outcomes Undermine Claims to Identification

Identification-oriented work generally purports to causally identify the 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 an average treatment effect ($ATE$). Following Manski (1995), the process of drawing 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 a 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 the quality of life measure is assessed, her relevant potential outcome for the quality of life measure is undefined. Standard experimental estimators of the $ATE$ (e.g., a difference-in-means) seek to estimate an undefined and thus unidentified quantity. Moreover, comparison of quality of life units that survive is not necessarily a principled experimental comparison because death may be endogenous to the treatment under study, undermining the treatment/control comparison.

More generally, an undefined variable 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, represent distinct scales. A deceased subject's quality of life is thus undefined. The difference in scales differentiates undefined outcomes from attrition or missingness. This distinction limits researchers' ability to productively apply strategies suggested to address attrition. Imputation methods make little sense when the onbserved values are measured on a distinct scale and resampling-based approaches do not address the underlying issue of distinct scales.[3]

I contend that the political 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 "cluster assigned," 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 four literatures that often make claims to causal identification: long-run devel-

[3]Bounding approaches on a distinct estimand, the survivor average causal effect (SACE) do resemble those used to bound interval estimates of estimands 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 |
| 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., Eggers, 2017), audit studies (Coppock, 2019; Slough, 2018), and policing (Knox, Lowe, and Mummolo, 2019).

opment, conflict, email audits, and incumbency advantage using experiments, natural experiments, regression discontinuity designs, or difference-in-difference strategies. Even if all standard identifying assumptions hold, if potential outcomes are undefined, the general quantities of interest, typically some $ATE$, $LATE$, or $ATT$, 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 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, the identification concern described here the fundamental concern is whether post-treatment actions alter the available set of strategies

or beliefs of a subsequent action or decision. To this end, theory introduces additional assumptions about the sequence of outcomes. I argue that for the purpose of identification, theory posits implications for what estimands could 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.

### 2.2 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, 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 of 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.

### 2.3 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 in 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," effectively 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 estimating broader "general equilibrium" effects which include selection. From a policy perspective, decisions based on the $SACE$ can be highly misleading (Joffe, 2011), 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 quite 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 particularly 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 on crime reporting by citizens and crime incidence.[4] Suppose that the campaign is cluster random assigned the "see something, say something" treatment messaging (flyers or targeted ads) to micro-neighborhoods within a city. Denote a binary treatment indicator, $Z_i$. They 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 $\mathcal{V}_i$.[5]

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

$$Y_i = \beta_0 + \beta_1 Z_i + \epsilon_i \tag{1}$$

The focus of enumerating the theory revolves around whether the estimator $\beta_1$ estimates the $ATE$ or any well-defined causal estimand. In general, I will calculate the estimand estimated by $\beta_1$ for each outcome for comparison to the analogous $ATE$ and $SACE$. I denote these quantities

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

[5]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.

[6]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.

$\Delta_{\mathcal{R}}$ and $\Delta_{\mathcal{V}}$.

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 across models. Collectively, these assumptions ensure comparability across both game theoretic and decision theoretic models. Further, among the game theoretic models, these assumptions ensure 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, $\lambda > 0$, drawn from pdf $f_\lambda(\cdot)$ and cdf $F_\lambda(\cdot)$ with support on the interval $[0, \bar{\lambda}]$. However, if a suspect that commits the crime is investigated, she pays a penalty $p > \bar{\lambda}$.

$B$ observes whether a crime occurs. If it does, 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.[7] If a crime is investigated, the bystander obtains a benefit, $\psi$, conceived as a taste for order or justice. These tastes vary across the population; $\psi \sim f_\psi(\cdot)$, where $f_\psi(\cdot)$ is a pdf

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

and $F_\psi(\cdot)$ is a cdf where $F_\psi(0) = 0$.

$O$ observes that a crime occurred and whether or not it was recorded. They choose to investigate or not to investigate. An investigation requires some allocation of effort by the officer at cost $\kappa$. $\kappa$ is a random variable drawn from pdf $f_\kappa(\cdot)$ with cdf $F_\kappa(\cdot)$ and support on $[0, \bar{\kappa}]$. While one can assume that officers earn a wage that satisfies their participation constraint, payment is not conditioned upon observable manifestations of effort. Instead, they face the threat of sanction, $s > \bar{\kappa}$ for failing to respond to crimes evaluated via random audit. Denote the expectation of a sanction for an audited officer, e.g. $s$ times the probability of sanction as $\alpha$. Assuming 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 three models. The extensive form of the full model (Case #4) appears in Figure 1. As is clear from Figure 1, 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 a unique subgame perfect Nash equilibrium 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 and 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. $\mathcal{R}_i = 1$

| **Case #1** | **Case #2** |
|---|---|
| *(1) A crime occurs with probability 1.* | *(1) With probability, $\rho$, 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 $\iota_R$. If a report is not received, nature investigates with probability $\iota_N$. | (3) If a report is received, nature investigates with probability $\iota_R$. If a report is not received, nature investigates with probability $\iota_N$. |
| (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 $\iota_R$. If a report is not received, nature investigates with probability $\iota_N$. | *(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.

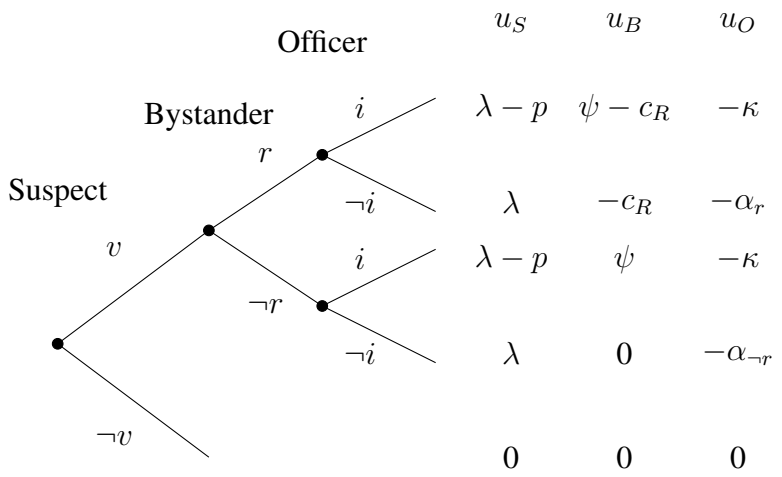



Figure 1: Extensive form representation for the Case #4.

if $v \cap r$ and that a case enters police records if it is investigated by police, i.e. $\mathcal{V}_i = 1$ if $v \cap i$. Second, estimands are expressed in terms of expectations over the potential outcomes of multiple units. While the equilibria characterized correspond to an equilibrium occurrence of reporting or investigation in one precinct, I examine differences in these outcomes in the aggregate (i.e., across precincts) between treatment and control.

**Case #1: Always Crime**

In the simplest variant of the model, there is always a crime that the bystander could report. Here, we are only concerned with the bystander's decision of whether to report or not. As shown in Appendix 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 in treatment are reduced relative to 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 "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* $\Delta_{\mathcal{V}} = ATE_{\mathcal{V}}$.

**Case #2: Exogenous Crime**

Case #2 parallels #1 except there exists exogenous selection into crime. With probability $\rho \in (0, 1)$, regardless of treatment assignment of the precinct, a crime occurs. Because there are

precincts with no crime, the bystander no longer faces the decision of whether or not to report where 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 $ATE$s in the first model.

However, with 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, $\rho$. With the present research design and the data described here, $\rho$ is not identifiable. Importantly, however, the difference-in-mean will maintain the same sign as the $SACE$.

**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 $ATE$s are identified; with such selection, the $ATE$s are neither defined nor identified, despite the fact that the experiment remains identical. These examples suggest 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 camplaign. Crime is committed when the surplus from committing the crime exceeds the expected probability of penalty. In this case, the campaign effects reporting through two channels. Conditional on a crime occuring, the lower net cost of reporting in treatment enlarges the set of bystanders (values of $\psi$) that would report. However, this also changes the suspect's calculus. She is less likely to commit the crime if it is more likely to be reported. These effects are counterveiling: treatment reduces crime rates (where there is no reporting) but increases reporting conditional on crime occuring. Without further assumptions on $f_\lambda$ or $f_\psi$, it is not possible to sign the resultant difference-in-mean estimate.

As in Case #2, selection into crime renders both $ATE$s 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 $\lambda$ denoted $\tilde{\lambda}$ where the suspect commits the crime in treatment (and thus also commits the crime in control). While the $SACE$ may be different from Case #2, depending on the joint distribution of $\lambda$ and $\psi$, it is positive. This occurs because the $SACE$ estimands effectively "close off" the crime (selection) channel.

**Remark 3.** *When crime is endogenous, 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:*

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

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

*Both expressions are ambiguous in sign.*

However, a naive difference-in-means estimate, does not recover $SACE_{\mathcal{R}}$ or $SACE_{\mathcal{V}}$ as shown in Remark 3. The ambiguous sign of this estimand 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 nature of 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 provides limited empirical leverage to identify any standard causal estimand.

**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 $ATE$s 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 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 utlities or information of the officer cannot solve the identification problems described here.

# 4 When is a Theory Necessary for Identification?

The hypothetical 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

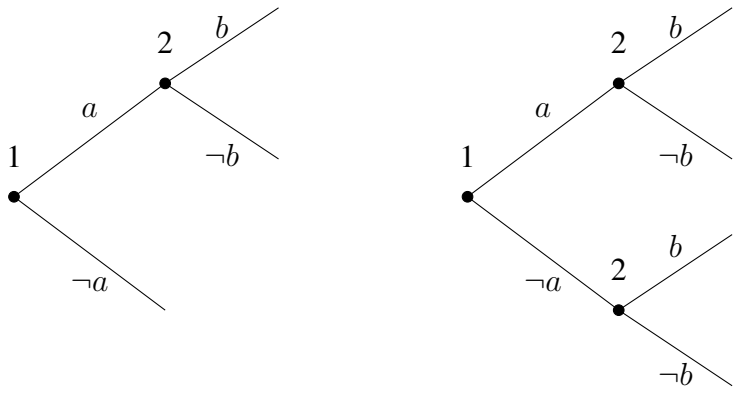
A feature of the models in Section 3 is that strategies are chosen sequentially, not simultaneously: the crime occurs (resp. does not occur), then the bystander reports or does not report it, then it is investigated (resp. not investigated). Given the emphasis on sequence, I restrict attention to dynamic 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, $H^{=}H^T$. 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 $h \in H \backslash H^T$.*

Symmetry in strategy sets is straightforward to visualize as a game tree. Figure 2 shows 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 $H^{\emptyset}$.

Examining the game trees in Figure 2, suppose 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 $E[a|Z = 1] - E[a|Z =$




ASYMMETRIC STRATEGY SETS

SYMMETRIC STRATEGY SETS

Figure 2: Extensive form representation of simple dynamic games with (right) and without (left) symmetry in strategy sets.

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 problem. Player 2 does not act if Player 1 chooses $\neg a$. With some 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 $ATE$ is 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 $SATE$, but the $ATE$ is not identified.

In contrast, in the right panel of Figure 2, 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 2. 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 2 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 $h$. Proposition 1 provides a general statement of this finding.

**Proposition 1.** *In an experiment in which standard identifying assumptions hold, if a 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 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 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 motivations of causal identification.

The primary threat to identification of the $ATE$ identified by Proposition 1 is indeed post-treatment selection. When this selection occurs in a sequence of post-treatment outcomes is criti-

cal. The $ATE$s of treatment on outcomes prior to and including the first instance of "selection" in 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.

Importantly, the invocation of a theory implies an increase in the amount, and possibly strength, of assumptions needed for identification. In addition to the standard empirical assumptions justifying the research design, we add assumptions about how actors behave and why to justify the design. In this sense, many existing claims of causal identification in the applied social science literature rely on the validity of an undelying (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.

### 4.2 When Should a Theory be Specified?

Proposition 1 implies that if a theory 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 theory of post-treatment behavior when outcomes are sequential.

Beyond the identification concerns outlined here, the invocation of a theory is in general 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 influences 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 even if crime is not changed by treatment. To that end, theory can be helpful even if not to justify identification of causal estimands.

To this point, I have focused on 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 strategy 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 (as opposed to individual actions), the specification of a theory allows for clear specification 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 *beliefs*? [to be completed.]

A natural extension considers other types of attitudinal outcomes. 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 in any way, attitudinal outcomes do not introduce the same threat of post-treatment selection as sequential behavioral outcomes. As such, theory to ground the elicitation of preferences may be less important than theory needed to ground behavioral outcomes.

# 5 Implications for Research Design

## 5.1 Best Practices for Experimental Design

[To be completed.]

## 5.2 Generalization from Experimental to Observational Designs for Causal Inference

To this point, I have focused on experiments and identification of the $ATE$. 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 information 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 an 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 decisions condition the set of available strategies, researchers should be particularly skeptical about how these causal processes reduce our ability to identify causal effects on long-run outcomes.

One implication of this variation is that the same "research design" can posit different degrees of need for a theory depending on what outcomes we seek to measure. Consider an experiment that is first used to identify immediate causal effects. Later, investigators seek to measure downstream outcomes. The argument here is that the same research design can require different assumptions to

identify causal effects on different outcomes. In general, the downstream analysis is apt to require a more "involved" theory than the original (immediate) analysis. The need for theory to ground identification increases in the possible complexity of post-treatment history.

Second, a researcher's ability to observe post-treatment behavior conditions the plausibility of assumptions imposed by a model. In an experimental intervention in which researchers design implementation of treatment and data collection, there is often room for observation – qualitative or quantitative – 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 a retrospective observational study. These observations can do much to ground a theoretical model.

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" (or shock in a natural experiment), the sequencing of interactions can be less self-evident.

In considering how these two dimensions overlap, there exists a dynamic the settings that most need theory to ground identification are precisely the settings in which we must rely upon the strongest/most assumptions. Figure 3 depicts this dynamic graphically. Several observations are noteworthy. In general, there is a positive correlation between the plausibility of standard empirical identifying assumptions and the plausibility of theoretical assumptions (to the extent that they are needed for identification). However, the plausibility of standard empirical assumptions does not vary across outcomes (or when we measure them) unless violations of SUTVA are somehow time/history dependent. Thus, the variation in the "need for theory" for a fixed research design (fixed $x$) is not a feature of other discussions of the a the research design's ability to identify causal effects.

This discussion also posits implications for the argument that "identification driven research"

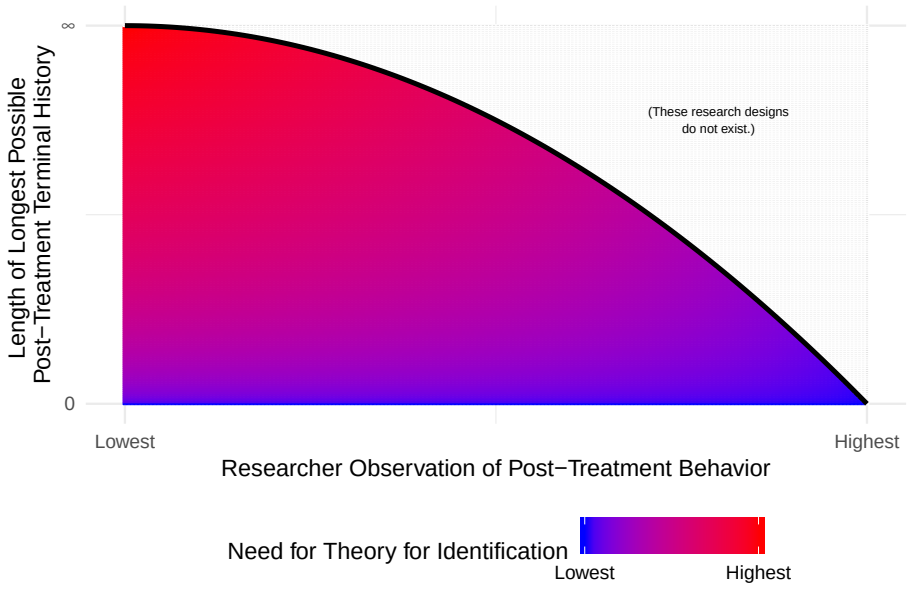



Figure 3: Graphical summary of the classification of research designs and the need for theory. The shape of the colored region that defines the "existance" of research designs is arbitrary.

(IDR) and theory are in tension (Huber, 2017).[8] In the framework developed in this paper, theory can be specified for any research design. While the practices advanced in this paper are not standard practice, consider the ramifications of a researcher's decision to focus efforts in areas where inference can be agnostic to theory. In this case researchers' efforts create the appearance of a tradeoff between theory and identification, when in the general case theory is necessary for identification.

## 6 Conclusion

In this paper I propose a role for theory in causal identification in reduced-form design-based research. The main innovation of the paper is to describe the need to address post-treatment selection in a strategic environment. The strategic environment in which much of social science is situated requires additional consideration of selection after treatment and its implications for identification of standard causal estimands. In so doing, I provide suggestions for why applied (substantive)

[8]This paper does not address Huber's (2017) claim that a focus on causal identification leads to the neglect of subjects in which the primary independent variable(s) are less exogenously manipulable by researchers or "nature."

theory helps to reason through problems of post-treatment selection, providing the assumptions necessary for identification. I argue that an explicit statement of a (dynamic) theory should be provided when at least some behavioral outcomes are sequential. Further, if the range of attitudinal measures elicited depend on "what happened" after treatment, a theory should be specified.

Ultimately, this paper seeks to provide advice for research design. Ideally empirical researchers should design research that allows for identification of main outcomes of interest. Moreover, theory provides insights into the interpretation of reduced-form causal estimands. Collectively, these considerations posit theory as an integral part of a minimal research design.

## References


Angrist, Joshua D., and J orn-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.

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.

Blair, Graeme, Jasper Cooper, Alexander Coppock, and Macartan Humphreys. 2016. "Declare-Design." R Package.

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.

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.

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. 2018. "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.

Knox, Dean, Will Lowe, and Jonathan Mummolo. 2019. "The Bias is Built In: How Administrative Records Mask Racially Biased Policing." Working paper available at `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3336338&download=yes`.

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.

Pearl, Judea. 2009. *Causality: Models, Reasoning, and Inference*. Second edition ed. New York, NY: Cambridge University Press.

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

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

Slough, Tara. 2018. "Bureaucrats Driving Inequality in Access: Experimental Evidence from Colombia." Working paper available at `http://taraslough.com/assets/pdf/JMP.pdf`.

Waltz, Kenneth N. 1979. *Theory of International Politics*. New York, NY: McGraw-Hill.

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:

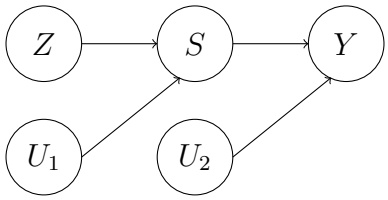



Figure 4: A graphical model depicting a causal process consistent with endogenous units of analysis if the potential outcome $Y(Z, S)$ is undefined for some units on the basis of the revelation of some $S(Z)$.

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 | $S(Z=1)$ | $S(Z=0)$ | $\bar{Y}(S=1, Z=1)$ | $\bar{Y}(S=1, Z=0)$ | $\bar{Y}(S=0, Z=1)$ | $\bar{Y}(S=0, Z=0)$ |
|---|---|---|---|---|---|---|---|
| **A**lways survivor | $\pi_A$ | 1 | 1 | $\bar{Y}_A(1,1)$ | $\bar{Y}_A(1,0)$ | - | - |
| If **T**reated survivor | $\pi_T$ | 1 | 0 | $\bar{Y}_T(1,1)$ | - | - | $[\bar{Y}_T(0,0)]$ |
| If **U**ntreated survivor | $\pi_U$ | 0 | 1 | - | $\bar{Y}_U(1,0)$ | $[\bar{Y}_U(0,1)]$ | - |
| **N**ever survivor | $\pi_N$ | 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:

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

However, because some of these quantities (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 $\psi$, the citizen will report if:

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

Given $F_\psi(\cdot)$, the cdf from which $\psi$ 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 $\rho$ proportion of cases in which there is crime, the $SACE$ follows from the calculation of the $ATE$ from Model B.1. Thus:

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

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 $E[U_S(\neg v)]$:

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

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 $ATE$s are undefined because some crimes do not occur. To compute the $SACE$, first it is useful to define two thresholds of $\lambda$ 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:

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

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 $E[U_S(\neg v)]$:

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

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

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

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

upon observing the crime, the bystander reports iff $\psi \geq \frac{c_R}{\iota_R-\iota_N}$; and upon recieving the report, the officer investigates iff $\kappa \leq \alpha_r$ but upon not receiving the report, the officer investigates iff $\kappa \leq \alpha_{\neg r}$. ■

# C Sketch of Proof for Proposition 1

Suppose that an experiment manipulates a single treatment $Z$. Assume that treatment assignment is ignorable and excludable and that SUTVA holds. First, suppose the extensive form of a game is not strategy set symmetric. Consider node $H^\emptyset$, representing the first post-treatment action in the model, $a$ as measured by $\mathcal{A}$. The $ATE$ can be written:

$$E[\mathcal{A}|Z=1]-E[\mathcal{A}|Z=0]$$

The ATE on outcome $\mathcal{A}$ is identified.

[Warning: this notation is rough.] For an arbitrary node, $h \in H \backslash H^T$ with a strategies $\{s_1,...,s_N\}$, denote the subsequent histories as $C$, where $C \subset H$ and each history takes the form of $\{h \cup s_n\}$. Suppose that $b$ that appears in the strategy set of at least one history $c \in C$ but not $\forall c \in C$. $b$ is measured by $\mathcal{B}$. The $ATE$ can be written:

$$\sum_{c\in C} Pr(c|Z=1)E[\mathcal{B}|Z=1,C=c]-\sum_{c\in C} Pr(c|Z=0)E[\mathcal{B}|Z=0,C=c] \tag{3}$$

But if $b$ is undefined as a strategy for some $c \in C$, $E[\mathcal{B}|Z=z,C=c]$ is undefined. As such the ATE is undefined and thus unidentified.

If the model exhibits strategy set symmetry, it must be the case that $b$ appears in the strategy set of all $c \in C$ if it appears in any strategy set $c \in C$. Thus all quantities in Equation 3 must be defined, and is identified under standard identifying assumptions. ■