Heterogeneous Treatment Effects and Causal Mechanisms*

Jiawei Fu† Tara Slough‡

June 21, 2023

Abstract

The credibility revolution advances the use of research designs that permit identification and estimation of causal effects. However, understanding which mechanisms produce measured causal effects remains a challenge. A dominant current approach to the quantitative evaluation of mechanisms relies on the detection of heterogeneous treatment effects with respect to pre-treatment covariates. This paper develops a framework to understand when such heterogeneous treatment effects can support inferences about the activation of a mechanism. We show first that this design does not provide evidence of mechanism activation without additional, generally implicit, assumptions. Further, even when these assumptions are satisfied, if a measured outcome is produced by a non-affine transformation of a directly-affected outcome of theoretical interest, heterogeneous treatment effects are not informative of mechanism activation. We provide new guidance for interpretation and research design in light of these findings.

*We thank Scott Abramson, seminar audiences at New York University, Princeton, and Berkeley, and participants at the NYU Abu Dhabi Theory in Methods Workshop for helpful feedback.

†Ph.D. Candidate, New York University. jf3739@nyu.edu

‡Assistant Professor, New York University. tara.slough@nyu.edu

The credibility revolution in empirical social science has motivated the largescale adoption of research designs that facilitate unbiased estimation of causal effects (Samii, 2016; Angrist and Pischke, 2010). Internally valid estimates of these effects allow for valid inferences about the causal effect of a treatment on an outcome. However, they do not, in general, provide evidence about *why* or *how* the treatment affected the outcome. These questions of why or how are ultimately questions about the activation and influence of causal mechanisms. Understanding the mechanisms through which causal effects are produced is central to our ability to use empirical evidence to understand social phenomena (Slough and Tyson, 2023*a*).

Applied researchers typically pursue a number of different approaches to ascertaining the mechanisms that generate causal effects. There exist at least four distinct approaches in the applied literature: (1) evaluation of the *sign* of treatment effects on a given outcome (e.g., Ashworth, Berry, and de Mesquita, 2023); (2) mediation analysis (Imai, Keele, and Tingley, 2010; Imai et al., 2011; Imai and Yamamoto, 2013); (3) multimethod research involving some form of qualitative or quantitative triangulation of causal findings (Levy Paluck, 2010; Dunning, 2012); and (4) the estimation of heterogeneous treatment effects (HTEs). The last approach—HTEs—involves measuring and comparing estimated treatment effects for various subgroups thought to be informative about mechanism activation. While this is currently the modal approach to (quantitative) mechanism-testing in political science, the theoretical properties of this approach are not well explored.

To examine the prevalence of the use of HTEs to infer causal mechanisms, we survey the 2021 volumes of three leading journals in political science: the *American Journal of Political Science* (AJPS), the *American Political Science Review* (APSR), and the *Journal of Politics* (JoP). We first identify the subset of papers that analyze quantitative empirical data. We then report the proportion of those empirical papers that use HTEs to make a claim about some mechanism(s). Finally, we report the proportion of papers using HTEs that interpret these quantities as providing information about causal mechanisms(s). Table 1 shows that in each of the three leading journals, a majority of quantitative studies estimate HTEs. Moreover, conditional on reporting any HTEs, the vast majority of articles (87% across three journals) interpret these quantities as providing information

Journal (Volume)	Number of articles: Total	Quant. empirical	Reporting HTEs	Pr(Report HTEs | Quant. empirical)	Pr(Mechanism test | Report HTE)
AJPS (65)	61	41	23	0.56	0.87
APSR (115)	106	75	40	0.53	0.90
JoP (83)	142	106	58	0.55	0.83
Total	309	222	121	0.55	0.87

Table 1: Authors' classification of articles published in three leading political science journals in 2021.

about mechanisms. Collectively, these figures indicate that almost half (48%) of recent quantitative empirical articles in these journals use HTEs to assess mechanisms.

Existing concerns about HTEs have largely focused on the *statistical* properties of relevant estimators and hypothesis tests. In particular, interaction effects are known to have low statistical power (e.g., McClelland and Judd, 1993). Moreover, estimation of HTEs with respect to many (pre-treatment) covariates risks multiple comparisons problems (Gerber and Green, 2012; Lee and Shaikh, 2014; Fink, McConnell, and Vollmer, 2014). While these criticisms are important, they are distinct from *theoretical* questions about how the presence or absence of HTEs links to causal mechanisms. We take on this challenge by asking: under what conditions do HTEs provide evidence of mechanism activation?

To answer this question, we develop a framework to formally link HTEs with respect to a covariate to the effect of a specific mechanism. To do so, we extend the workhorse causal mediation framework (Imai, Keele, and Tingley, 2010). Our analysis probes what can be inferred about the indirect effect of a specific mechanism from HTEs with respect to the covariate. Our results characterize the conditions under which heterogeneity in conditional average treatment effects (CATEs) is a sufficient condition to show that there exists at least one unit for which the indirect effect of the mechanism of interest is non-zero. A mechanism is *active* when its indirect effect is non-zero for some unit.

We first show that two classes of exclusion assumptions are required to link HTEs to a specified mechanism. These assumptions hold that the covariate of interest is excluded with respect to (1) the average indirect of other mechanisms and (2) the average direct effect of the treatment

on the outcome. In the absence of these assumptions, the relationship between HTEs and the indirect effect of specified mechanism is unspecified. In this sense, these assumptions are implicitly invoked by current practice, but are generally not stated or defended.

When these exclusion assumptions hold, what can we learn about a mechanism of interest from HTEs with respect to a given covariate? Our main results characterize what we can learn from the existence or non-existence of HTEs. First, consider the case in which a measured outcome is *directly* affected by the mechanism of interest. We show that under these assumptions, the existence of HTEs provides evidence of mechanism activation. This broadly conforms to current practices in applied research, albeit while making the underlying assumptions explicit. However, when there do not exist HTEs, there are two possible explanations: (1) the indirect effect effect of a mechanism is not moderated by any covariate, measured or unmeasured; and/or (2) the posited relationship between the covariate and the indirect effect of the mechanism was mis-specified (i.e., the theory is wrong). Neither possibility distinguishes between an active or inactive mechanism. This means that a lack of HTEs cannot "rule out" a mechanism by showing that it is inactive. This finding contradicts standard interpretations of (the absence of) HTEs in the applied literature.

We further consider the common case in which we observe outcomes that are *indirectly* affected by a mechanism. For example, in a political economy model, a mechanism may change an actor's utility, but researchers observe only the actor's discrete choice/behavior. In political psychology, a treatment may affect a subject's latent attitudes, but researchers observe their survey response on a Likert scale. When observed outcomes are generated by a non-affine mapping of the directly-affected outcome (as in both of these examples), the relationship between HTEs and mechanisms changes. In this case, the existence of HTEs with respect to a covariate no longer provides evidence of mechanism activation, even when both exclusion assumptions hold. The logic for this result is straightforward: the non-linear mapping of the mechanism's influence into an observed outcome breaks the additive separability of the indirect effect of interest from other indirect and direct effects, which undermines our ability to link HTEs to the indirect effect of a mechanism.

This paper makes three principal contributions. First, we contribute to literature on varying uses

of HTEs. Our focus, in line with most current empirical applications, is on the use of HTEs to learn about the mechanisms that underlie treatment effects. We argue that it is important to distinguish the use of HTEs to learn about mechanisms from the use of HTEs for extrapolation, prediction, or targeting of treatment. The latter class of concerns—extrapolation, prediction, and targeting—have been more dominant in the recent methodological literature. Recent articles advocate extrapolation from experimental treatment effect estimates to treatment effects in a target population (Egami and Hartman, 2022; Devaux and Egami, 2022). Other recent contributions suggest exploiting (estimated) treatment effect heterogeneity to better target treatments in the future (Kitagawa and Tetenov, 2018; Athey and Wager, 2021). These methods are aided by the use of machine learning approaches to the detection of heterogeneity (e.g., Grimmer, Messing, and Westwood, 2017; Athey, Tibshirani, and Wager, 2019). In contrast, our results suggest that learning about why we observe causal heterogeneity relies on a deductive theoretical mapping of covariates to mechanisms, which is unlikely to be aided by machine learning approaches.

Second, we expand a growing literature on the theoretical implications of empirical models (TIEM) (Bueno de Mesquita and Tyson, 2020; Ashworth, Berry, and Bueno de Mesquita, 2021; Ashworth, Berry, and de Mesquita, 2023; Abramson, Koçak, and Magazinnik, 2022; Slough, 2022). We make two central interventions to this literature. First, our framework makes explicit links between a causal mediation framework that is used more prominently by empiricists and theoretical models employed in formal theory. Second, we introduce questions about how measured outcomes relate to theoretical constructs by distinguishing between directly- and indirectly-affected outcomes. While measurement is central to recent TIEM work on evidence accumulation in a cross-study environment (Slough and Tyson, 2023*a*,*b*,*c*), it has not been widely explored in the single-study environment.

Finally, we provide practical guidance for empirical researchers who want to learn about which mechanisms generate observed effects. Our assumptions and results reveal a minimal set of attributes of an applied theory that can support the use of HTEs to learn about mechanisms. Two of these attributes, the relationships between (1) a covariate of interest and other mechanisms and (2)

measured outcomes and theoretical objects of interest, are generally absent from applied work in current practice. Second, we show how interpretation of HTEs can be improved, returning to the statistical problems that are well known in this literature. Third, we discuss how our analysis can be used to inform prospective research design. Finally, we show that in order to infer mechanism activation from HTEs in the case of indirectly-observed outcomes, stronger theoretical and untestable empirical assumptions are generally necessary. Collectively, these suggestions allow practitioners to accurately use—or, when indicated, avoid—HTEs as a quantitative test of mechanisms.

1 Current Practice

As reported in Table 1, 87% of the articles that report HTEs interpret these quantities as tests of a mechanism. Two interpretations of HTEs are common. First, the *presence* of HTEs with respect to a specific covariate provides evidence that a mechanism is active. For example, Malis and Smith (2021) study the effect of US presidential visits and foreign leader visits to the US on foreign leader removal. They argue that in-person diplomacy with the US provides a credible signal that the US believes the foreign leader is likely to remain in office. This, in turn, deters doemestic competitors from challenging the incumbent, increasing the incumbent's tenure in office. To assess this mechanism, they show that the effect of in-person diplomacy with the US on leader survival is *stronger* in more unstable countries where in-person diplomacy provides greater information about the strength of the foreign leader. Here, the presence of HTE in a measure of instability of the recipient country is interpreted to support the activation of the proposed informational channel.

Second, the *absence* of HTEs with respect to a given covariate is frequently used to "rule out" the activation of a possible mechanism (often called an alternative explanation). For example, Moscowitz (2021) argues that the proportion of in-state residents in one's local media market increases rates of voter political knowledge and split-ticket voting in the US, countering trends toward the nationalization of politics. The paper provides evidence that news coverage is the mechanism that drives this effect. However, it also seeks to rule out an alternative mechanism that holds that more in-state residents lead to more campaign advertising about in-state candidates,

which in turn increases voter knowledge. Since advertisements generally air when incumbents are contesting re-election (but not otherwise), Moscowitz (2021) codes an indicator that takes the value of "1" when the survey was fielded during an election season (when the incumbent was running) and "0" otherwise. The effect of the share of in-state residents in a local media market does not detectably vary when the incumbent is running versus not running for re-election. This *lack* of heterogeneity is used to provide evidence against the advertisement-based mechanism.

Both of the above examples are exceptionally clear in delineating the mechanisms under consideration using HTEs, and thereby serve as exemplars of current practice. Our concern in this paper is that current practice with respect to HTEs and the detection of causal mechanisms can mislead. We proceed from these empirical examples to a purely theoretical motivating example to illustrate the concern.

2 Motivating Example: Exogenous Shocks and Voting

Consider a large class of natural experiments on the effect of some exogenous shock, denoted ω, on voter beliefs and behavior in a democracy.[1] This shock could be a natural disaster (e.g. Healy and Malhotra, 2010; Achen and Bartels, 2017), a disease outbreak (e.g., Baccini, Brodeur, and Weymouth, 2021), an economic crisis (e.g. Wolfers, 2002), or even a seemingly-irrelevant event (e.g. Healy, Malhotra, and Mo, 2010). In order to characterize the effect of the shock on voter beliefs and behavior, we adapt a formal model by Ashworth, Bueno de Mesquita, and Friedenberg (2018).

We will assume that an incumbent at the time of the shock is of type $\theta \in \{\underline{\theta}, \overline{\theta}\}$, where $\overline{\theta} > \underline{\theta}$, such that a politician of type $\overline{\theta}$ is "good type" and a politician of type $\underline{\theta}$ is a "bad type." Voters do not observe the politician's type directly, but may be able to learn about their type from observed governance outcomes. The governance outcome is given by:

$$g = f(\theta, \omega) + \varepsilon. \tag{1}$$

[1] One could alternatively conduct a survey or field experiment in which researchers provide information on an exogenous shock that is not otherwise observed by voters.

In this formulation, higher values of ω correspond to a more adverse shock (e.g., the intensity of a natural disaster). Function f is monotonically increasing in θ and decreasing in ω. ε is an idiosyncratic shock to the governance outcome that is drawn from a symmetric, continuously differentiable probability density function, ϕ, that satisfies the monotone likelihood ratio property relative to g.[2]

Each voter's utility from a politician depends on the politician's type, θ, and a valence shock for the incumbent, v_i. Here, valence captures any attribute of the incumbent that does not depend on their type, including but not limited to bias toward a candidate or ideological closeness. Politician type and valence are additively separable, such that voter utility is given by:

$$u_i = \begin{cases} \theta^I + v_i & \text{for the incumbent } (I) \\ \theta^C & \text{for the challenger } (C) \end{cases} \tag{2}$$

In the population, $v_i \sim U(-1, 1)$. Voters have heterogeneous prior beliefs about the probability that incumbent is of type $\overline{\theta}$, formally $\pi_i(\overline{\theta}) \in [0, 1]$, where $\pi_i(\overline{\theta}) \sim F_\pi$. For simplicity, we assume that voters share a common prior $\pi^C \in (0, 1)$ for the challenger.[3] Voters vote either for the incumbent or challenger. The sequence is as follows:

1. Nature reveals shock ω and voters observe both the shock and the governance outcome.
2. Voters update their beliefs about the incumbent's type.
3. Voters vote for either the incumbent or the challenger.

Posterior beliefs and voting behavior are straightforward to characterize. Given a shock, ω and a voter's prior, π_i, a voter's posterior belief about the incumbent's type, θ^I, is given by:

$$\beta(\overline{\theta}|\pi_i, \omega) = \frac{\pi_i \phi(g - f(\overline{\theta}, \omega))}{\pi_i \phi(g - f(\overline{\theta}, \omega)) + (1 - \pi_i)\phi(g - f(\underline{\theta}, \omega))} = \frac{1}{1 + \frac{1-\pi_i}{\pi_i} \frac{\phi(g - f(\underline{\theta}, \omega))}{\phi(g - f(\overline{\theta}, \omega))}} \tag{3}$$

[2]Formally, this implies that if $x' > x$, then $\frac{\phi(g-x')}{\phi(g-x)}$ is strictly increasing in g.

[3]The superscript C denotes the challenger. Any belief without a superscript pertains to the incumbent.

A voter will vote for the incumbent if and only if:

$$\beta(\overline{\theta}|\pi_i, \omega) + v_i \geq \pi^C \tag{4}$$

2.1 From Theory to Empirical Research Design

Mapping this model onto the empirical research design, we will assume that $\omega \in \{\omega', \omega''\}$ denotes a binary treatment, where ω' indicates no exposure to the shock and ω'' denotes exposure to the shock. The first outcome, $y_1(\omega)$, measures voters' expected utility from the incumbent. It is obviously difficult and rare to measure utility directly, though one could, in principle, elicit willingness-to-pay. A voter's expected utility from a vote for the incumbent is given by:

$$y_{i1}(\omega) = \beta(\overline{\theta}|\pi_i, \omega) + v_i \tag{5}$$

The second outcome, y_2, measures each voter's vote choice for the incumbent. Vote choice is obviously a more standard outcome in literature on voter behavior. This outcome is given by:

$$y_{i2}(\omega) = \begin{cases} 1 & \text{if } \beta(\overline{\theta}|\pi_i, \omega) + v_i - \pi^C \geq 0 \\ 0 & \text{else} \end{cases} \tag{6}$$

Our model can be represented as a directed acyclic graph (DAG), as depicted in Figure 1. This representation clarifies a number of assumptions of the model. First, the mechanism through which the shock affects voter preferences and behavior is through voter learning. This is evident because the only path from the shock (ω) to the outcomes $y_1(\omega)$ and $y_2(\omega)$ passes through the voter's posterior beliefs about the incumbent's type. Voter learning is moderated by the voter's prior belief about the incumbent, $\pi_i(\overline{\theta})$. Throughout this paper, we will indicate causal moderation through the arrow pointing to a path rather than a node.[4] For the purposes of exposition, the mediator, posterior

[4]We could define an extra node to indicate this interaction. We do not do so in the interest of parsimony. Note that the representation of "interaction" effects in DAGs is not standardized (Nilsson et al., 2021).

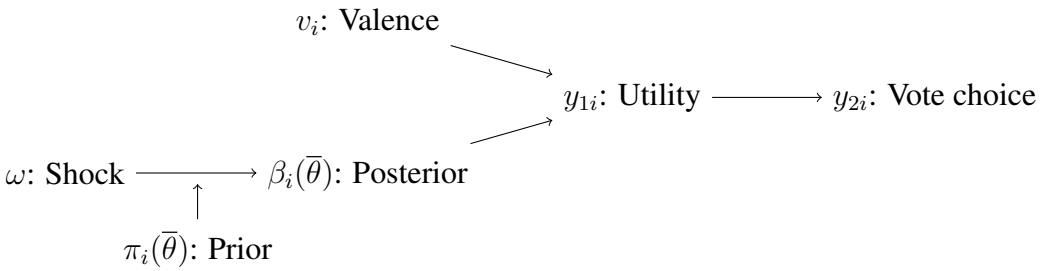

Figure 1: Directed acyclic graph representation of the model of voter preferences and behavior. The arrow from the prior to the path between the shock and the posterior indicates that the prior moderates voter updating on the incumbent, as in (3).

beliefs, is unobserved to researchers. We will assume, however, that researchers have a measure of π_i and v_i from a baseline survey.

Empirical researchers typically will not fully understand the causal structure represented in Figure 1. If this were the case, a researcher may mistakenly think that a shock affects assessments of valence, such that v_i is a moderator. (For reference, we depict this alternative DAG that is inconsistent with our model in Figure A1.) In the absence of measured mediators, the researcher could assess the mechanisms using by evaluating differences in conditional (or subgroup) treatment effects. Specifically, we will follow common practice by supposing that researchers estimate conditional average treatment effects (CATEs) at different levels of a moderator, X, as follows:

$$CATE(y, X) = E[y(\omega = \omega'') - y(\omega = \omega')|X = x] \tag{7}$$

We will say that treatment effects are heterogeneous if for some $x, x' \in X$ where $x \neq x'$, $CATE(y, X = x) - CATE(y, X = x') \neq 0$. When used to detect mechanisms, researchers typically assert that this form of heterogeneity gives evidence of mechanism activation or presence. Indeed, this was the structure of the claims about HTEs and mechanisms in both empirical examples.

Under our model of voter updating and behavior, do heterogeneous treatment effects provide evidence that the relevant mechanism is voter learning? *Ex-ante*, empirical researchers do not know that learning is the active (or operative) mechanism. To learn about mechanisms, many researchers

will estimate heterogeneous treatment effects, typically pointing to heterogeneity as evidence of mechanism activation. When is this approach valid? When does it yield invalid substantive inferences about mechanisms? To develop intuitions, we evaluate four heterogeneous treatment effect combinations using moderators $\mathbf{X} = \{\Pi, V\}$, where Π is the set of all possible values of π_i and V is the set of all possible values of v_i, and outcomes $y \in \{y_1, y_2\}$. Remark 1 shows that for the outcome measuring a voter's expected utility from a vote for the incumbent, detecting heterogeneity in CATEs correctly provides evidence that the mechanism is voter learning about the incumbent's type.

Remark 1. *For the outcome measuring voter preferences,* y_1*,*

(a) Given $\pi \neq \pi' \in \Pi$*,* $CATE(y_1, X = \pi) - CATE(y_1, X = \pi') \neq 0$*.*

(b) Given $v \neq v' \in V$*,* $CATE(y_1, X = v) - CATE(y_1, X = v') = 0$*.*

(c) If $CATE(y_1, X = x) - CATE(y_1, X' = x') \neq 0$*, then* $x, x' \in \Pi$*.*

(All proofs in appendix.)

Researchers will detect heterogeneity with moderator π (by a). Moreover, they will only detect heterogeneity in π, not in v, (by c) supporting the inference that the learning mechanism produces the observed ATE. In contrast, HTEs are not observed for the (non)-moderator v (by b). Here, researchers are unlikely to mis-attribute the mechanism through analysis of HTEs.

However, for the vote choice for the incumbent, y_2, the results from Remark 1 change. First, researchers may observe HTEs for different levels of valence, v, as well as for different levels of prior beliefs, π. This would lead most researchers to infer that the effect of the shock *does* work through some channel involving valence in addition to a channel involving voter learning.

Remark 2. *For the outcome measuring voter choice* y_2*,*

(a) Given $\pi \neq \pi' \in \Pi$*,* $CATE(y_2, \pi) - CATE(y_2, \pi') \neq 0$ *almost everywhere.*

(b) Given $v \neq v' \in V$*,* $CATE(y_1, v) - CATE(y_1, v') \neq 0$ *almost everywhere if* $\min\{v, v'\} < \pi^C$*.*

(c) If $CATE(y_1, x) - CATE(y_1, x') \neq 0$*, then* $x, x' \in \Pi$ *or* $x, x' \in V$*.*

Why do we see HTEs in v for vote choice? Recall that each voter votes for the incumbent if and only if $\beta(\overline{\theta}|\pi_i, \omega) + v_i - \pi^C \geq 0$. But this binary choice means that voter beliefs and valence are no longer additively separable with respect to the discrete outcome, vote choice. As such, with a sufficiently large sample size (and thus sufficient statistical power), researchers are apt to detect HTEs even when a mechanism is not active. Using standard interpretations of these tests, this leads to Type-I errors in our inferences about mechanism activation.

This example yields three important observations that we develop by proposing a new framework:

1. The use of HTEs does, in some cases (i.e., Remark 1), provide information about mechanism activation. This accords with current practice.

2. The use of HTEs to measure mechanism activation relies on assumptions about the relationship between moderators and mediators of interest which are typically implicit.

3. The contrast between Remarks 1 and 2 in which the theory (and thus mechanism) is fixed but outcomes differ hints that the use of HTEs for assessing mechanism activation depends on the data-generating process behind the outcome of interest.

3 Framework

We introduce a framework that we use to analyze the relationship between HTEs—as estimated by treatment-covariate interactions—and the detection of causal mechanisms. Our framework is built upon the potential outcomes framework or Neyman-Rubin causal model (Neyman, 1923; Rubin, 1974). We denote treatment by Z and potential outcomes by $Y(Z)$. In order to consider HTEs with respect to pre-treatment moderators, denote the set of *measured* pre-treatment covariates by $\mathbf{X}$. Clearly, it need not (and generally will not) be the case that all variables in $\mathbf{X}$ moderate the effect of Z on Y.

We will further denote a causal mediator by M. A valid mediator should: (1) be affected by treatment, Z, and (2) have a non-zero effect on the outcome Y. Importantly, a causal mediator

could be affected by some covariate(s) in **X**. This "arrow" from some X to the value of the mediator is essential for using heterogeneous treatment effects to detect causal mechanisms. In the above example, the mediator—posterior beliefs—is affected by both treatment and prior beliefs. We use $M(Z, X)$ to denote the potential outcomes of causal mediator M given treatment Z and covariates X.

3.1 Causal Effects

In order to understand when HTEs can allow for detection of mechanisms, we formalize existing informal conventions. To do so, we decompose the total effect (on a unit, i) into direct and indirect effects, as is standard in the causal mediation literature (Imai and Yamamoto, 2013). Suppose that there exist $J \geq 1$ mediators (or mechanisms), indexed by $M_1, ..., M_J$. Without loss of generality, our goal is to understand whether the effect is mediated by M_1. Given two treatment values, $z, z' \in Z$, the total effect of Z on Y is:

$$TE(z, z'; X) = Y(z, M_1(z, X), ..., M_J(z, X), X) - Y(z', M_1(z', X), ..., M_J(z', X), X) \quad (8)$$

Our notation varies slightly from conventional presentations of mediation that only consider two mechanisms (mediators) (Imai, Keele, and Tingley, 2010). To this end, $j-$ and $j+$ denote index $h \in J$ such that $h < j$ and $h > j$ respectively. Treatment effects may consist of direct (DE) and indirect (IE_j) effects, as follows:[5]

$$DE(z, z'; X) = Y(z, M_1(z, X), ..., M_J(z, X), X) - Y(z', M_1(z, X), ..., M_J(z, X), X) \quad (9)$$

$$\begin{aligned} IE_j(z, z'; X) = Y(z', M_{j-}(z', X), M_j(z, X), M_{j+}(z, X), X) - \\ Y(z', M_{j-}(z', X), M_j(z', X), M_{j+}(z, X), X) \end{aligned} \quad (10)$$

The direct effect, $DE(z, z'; X)$ represents the direct effect of Z on Y holding mediators at

[5]The intuition for our notation is as follows: define $IE_0(z, z'; X) = DE(z, z'; X)$. This implies that the first term of $IE_j(z, z'; X)$ and the second term of $IE_{j-1}(z, z'; X)$ cancel out.

potential outcomes $M_j(z, X)$. It is not necessary to believe that treatments produce unmediated (direct) effects on outcomes to use this framework. We allow for direct effects in the interest of generality. The indirect effects, measures the effect on the outcome that operates by changing the potential outcome of the mediator. As is standard, we can then re-write the total effect as follows:

$$TE(z, z'; X) = DE(z, z'; X) + \sum_{j=1}^{J} IE_j(z, z'; X), \tag{11}$$

which is defined at the unit, or individual level. If we evaluate expectations over $\neg X$, we obtain:

$$\begin{aligned} ATE(z, z') &= E_{\neg X}[Y(z) - Y(z')] && (12)\\ &= E_{\neg X}[DE(z, z'; X) + \sum_{j=1}^{J} IE_j(z, z'; X)] && (13)\\ &= ADE(z, z'; X) + \sum_{j=1}^{J} AIE_j(z, z'; X) && (14) \end{aligned}$$

We use ADE and AIE to denote average direct effect and average indirect effect. Throughout the paper, we assume this expectation is well-defined. Our framework proceeds by linking the indirect effect associated with a mechanism to heterogeneous treatment effects with respect to a covariate. To this end, define conditional average treatment effects as follows:

Definition 1 (Conditional Average Treatment Effect)**.** *Consider pre-treatment covariate $X_k \in \boldsymbol{X}$. Given that $z \neq z' \in Z$, the conditional average treatment effect (CATE) with respect to $X_k = x$ is:*

$$CATE_Y(X_k = x) = E_{X_{\neg_k}}[Y(Z = z) - Y(Z = z')|X_k = x]$$

There are many ways to define and measure treatment effect heterogeneity. In this article, we consider HTEs with respect to pre-treatment moderators. This adheres to the common practice of using HTEs to detect mechanisms that we documented in Table 1.

Definition 2 (Heterogeneous treatment effects)**.** *HTEs exist with respect to pre-treatment covariate*

$X_k \in X$ *if* $CATE_Y(X_k = x) \neq CATE_Y(X_k = x')$ *for some* $x \neq x' \in X_k$.

With this framework, it is now possible to express the research question more precisely. First, recall our research question: "under what conditions do HTEs with respect to pre-treatment covariates provide evidence of mechanism activation?" By mechanism activation, we mean that the indirect effect of a mechanism j is non-zero for some unit. The use of heterogeneous treatment effects registers the expectation that the mechanism need not be active for all units in the population. Our question then can be stated more precisely: "Under what conditions are HTEs with respect to a covariate X_k sufficient to show that there exists some unit for which $IE_j(z, z'; X) \neq 0$?"

3.2 Mechanisms and Outcomes

In standard discussions of research design, researchers often focus immediately on treatment effects on *measured* outcomes. Yet, as our example suggests, mechanisms may instead act directly upon *latent*, or unobserved outcomes. Literature on external shocks and voting behavior typically emphasizes vote choice for the incumbent candidate or party as the primary outcome measure. Yet, under our theoretical model (and many in the literature), the shock affects a voter's utility from a vote for the incumbent via their beliefs about the incumbent's type. So the mechanism—voter learning—operates directly on voter utility. The standard outcome that is observed—vote choice—is a (non-linear) mapping of that directly-affected outcome of interest.

This distinction between an underlying unmeasured variable—utility—and observed outcomes is standard in the formal theory literature. But this feature is much more widespread. For example, in studies that measure effects of various interventions on attitudes, attitudes are generally viewed as a latent variable that is affected by treatment. We observe responses to survey questions which are some mapping of latent attitudes. For example, in a discussion of attitude formation, Coppock (2022: p. 43-44) writes: "Respondents combine their considerations via a quasi-random process into a latent attitude, which I'll call y_i^* ... A latent attitude is translated into a measured attitude via a survey question $Q : y_i = Q(y_i^*)$."

In mechanistic analysis, this distinction between *directly* and *indirectly* affected outcomes merits consideration. We formalize the distinction in Definition 3, while providing concrete examples

Outcome type	Notation	Examples
Directly affected by mechanism	Y	Utility, attitudes
Indirectly affected by mechanism	$h(Y)$	Behavior/choice, survey responses

Table 2: We distinguish between outcomes that are directly or indirectly affected by a mechanism.

in Table 2. The distinction between unobserved (latent) and observed outcomes is well-known in literature on measurement models (Poole and Rosenthal, 1985; Fariss, Kenwick, and Reuning, 2020). Moreover, unobserved outcomes—like utility or attitudes—are widespread, if implicit, in most theories, whether formal or informal. Despite these commonalities in the empirical and theoretical literatures, insufficient attention has been devoted to the implications of these issues of outcome measurement for the study of causal mechanisms. Our analysis below shows that this distinction is consequential for what can be learned about mechanisms from HTE.

Definition 3. *Given treatment Z, let variable $\widetilde{Y}(Z)$ be a (potential) outcome. If*

1. *There exists another variable $Y(Z)$ that causally precedes $\widetilde{Y}$; and*

2. *$\widetilde{Y} = h(Y)$ where $h(\cdot)$ is not an affine function;*

*then we call $\widetilde{Y}$ an **indirectly-affected outcome**. Otherwise, $\widetilde{Y}$ is a **directly-affected outcome**.*

Ultimately determining whether an outcome is directly or indirectly affected by a mechanism is a *theoretical* commitment. But this theoretical commitment will have important implications for how we assess mechanisms empirically. Specifically, where this commitment is not made explicit, either through a theoretical model or clear verbal argumentation, we will show that HTEs cannot provide evidence of mechanism activation.

The mediation framework that we build upon makes no distinctions based on the type or causal sequencing of outcomes. Yet, our motivating example reveals a distinction in what we can learn about mechanisms from HTEs on expected utility (the directly-affected outcome) versus vote choice (the indirectly affected outcome). This suggests that, in at least some cases, more structure is necessary to bridge the disjuncture between the mediation framework and many applied theories. This distinction between direct- and indirectly-affected outcomes is our solution to this

disjuncture; there are certainly other ways to impose sufficient structure to bridge theoretical models and the statistical mediation framework. Our simple classification of outcomes proves quite useful for our analysis of what can be learned from HTEs.

4 HTEs and Mechanisms

When do heterogeneous treatment effects with respect to some covariate, X_k, provide evidence that a mechanism is active? To answer this question, we must first operationalize the concept of a theoretical mechanism. We view mediators—whether measured or unmeasured—as representations of a theoretical mechanism. For example, in the earlier example about exogenous shocks and voter behavior, the voter's posterior represents the mediator. If a mediator is measured, researchers can, in theory, use off-the-shelf estimators to estimate the direct and indirect effects described above. Yet, there are two substantial limitations of mediation analysis. First, researchers may not have the ability to measure mediators. Not all mechanisms are measurable and even fewer are measured. Second, mediation analysis typically relies upon an additional ignorability assumption for identification. In our notation, this ignorability assumption, $Y_i(z, m; X) \perp M_i(z, X) \mid Z_i, X_i$, holds that conditional on treatment and pre-treatment covariates X_i, potential outcomes are independent of the potential outcomes of the mediator (Imai, Keele, and Tingley, 2010). Critics allege that this assumption is unlikely to obtain and show that failure of this assumption generates bias in estimates of indirect effects (Gerber and Green, 2012).

The mediation framework allows us to precisely characterize HTEs with respect to covariates. However, it does not yet provide enough structure to link HTEs (or lack thereof) to mechanisms. To do so, we develop the concept of a mechanism indicator variable (MIV) and impose two assumptions. A MIV for a given mechanism induces a differential (average) causal effects through the mechanism of interest. This can be expressed in terms of (average) indirect effects. To economize notation, we will denote $AIE_j(X_k = x) = AIE_j(z, z'; X_k = x)$ as the average indirect effect of mechanism (mediator) j when $X_k = x$.

Definition 4 (Mechanism indicator variable (MIV))**.** *A pre-treatment covariate X_k is a mechanism*

Figure 2: The two panels depict $X_k \in \mathbf{X}^{MIV}$ for mechanism M_1 graphically. Either panel is consistent with Definition 4.

indicator variable for mechanism j if for some $x, x' \in X_k$, $AIE_j(X_k = x) \neq AIE_j(X_k = x')$.

We then denote $\mathbf{X}^{MIV}$ as the (possibly empty) set of measured covariates that satisfy Definition 4. Intuitively, if $X_k \in \mathbf{X}^{MIV}$, then covariate X_k can serve as an indicator for a mechanism/mediator of interest.[6] Under our definition of MIVs, it could be the case that X_k moderates the effect of the treatment on the mediator (M_j). Interestingly, it could also be the case that X_k moderates the effect of the mediator on the outcome. Both possibilities are depicted in Figure 2. The detection of mechanisms using treatment-by-covariate interactions requires researchers to postulate a MIV for a given mechanism.

However, postulating a MIV is not sufficient to use HTEs to detect the activation of a mechanism. Specifically, we must also be concerned with whether a given covariate is a MIV for *other* mediators in addition to M_j. Specifically, in order to do this, researchers need two additional exclusion restrictions which limit the number of mediators that are affected by a given covariate, X_k. Logically, if a single covariate moderates multiple mechanisms, we cannot use heterogeneity in that covariate in order to isolate our mechanism of interest (M_j). Assumption 1 rules out the possibility that direct effects depend on X_k and Assumption 2 rules out the possibility that the indirect effects produced by other mechanisms are moderated by X_k. Importantly, these assumptions do not rule out a direct path from X_k to the outcome Y. Nor do they rule out a direct path between X_k and any other moderator, $M_{\neg j}$, so long as X_j does not moderate the effect of treatment through that mechanism.

Assumption 1 (Exclusion I)**.** *Given $z, z' \in Z$ and $x, x' \in X_k$, X_k is excluded to the direct effect*

[6] A slightly stronger version of Definition 4 holds when Y is continuously differentiable with respect to M_j and Z . In this case, Definition 4 can be expressed as $\frac{\partial}{\partial X_k}\left(\frac{\partial Y}{\partial M_j}\frac{\partial M_j}{\partial Z}\right) \neq 0$.

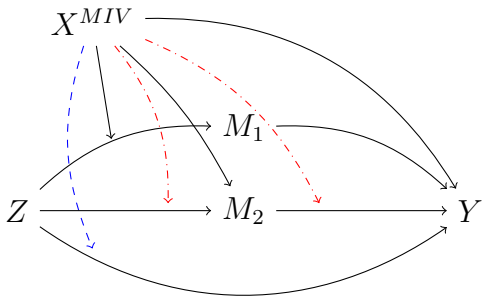

Figure 3: Assumption 1 rules out the blue dashed path. Assumption 2 rules out both of the red dot-dashed paths. All black solid paths are permissible under Assumptions 1 and 2.

such that $ADE(z, z'; X_k = x) = ADE(z, z'; X = x')$.

Assumption 2 (Exclusion II)**.** *Given* $z, z' \in Z$ *and* $x, x' \in X_k$*,* X_k *is excluded to the indirect effect of any other mechanism,* $j' \neq j$*,* $IE_{j'}$*, if:* $AIE_{j'}(x) = AIE_{j'}(x')$.

Assumptions 1-2 constrain the relationship between a moderator, X_k, and other mechanisms/the direct effect. Figure 3 depicts violations of Assumptions 1-2 graphically for a MIV (X^{MIV}) of mechanism 1, M_1. Neither assumption rules out a direct causal relationship between X_k and outcome Y. In Figure 3, such a relationship is depicted in the arrow from X_k to Y. Assumption 1 rules out the blue dashed path from X^{MIV} to the direct effect of Z on Y. Assumption 2 rules out either/both red dot-dashed path from X^{MIV} to the effect of Z on a different mechanism, M_2. This means that X_k does not moderate the effect of Z on M_2. Nor does X_k moderate the effect of M_2 on Y. However, there can be a direct arrow from X_k to M_2. This means that X_k can cause (or predict) the level of M_2, but it cannot interact with treatment in any way. These assumptions are crucial for the use of HTEs to detect mechanisms. In their absence, we cannot link a heterogeneous treatment effect to a unique theoretical mechanism representation.

Use of HTEs in order to learn about mechanism activation represents one alternative to mediation analysis. Assumptions 1-2 form the core assumptions underpinning this use of HTE. The comparison to mediation invites a comparison of these exclusion assumptions to the assumption of sequential ignorability in mediation analysis. It is useful to note that there is no logical ordering of the two types of assumptions: the exclusion assumptions do not imply sequential ignorability, nor

does sequential ignorability imply the exclusion assumptions. This means that HTEs cannot said to be a "more agnostic" or "less agnostic" quantitative test of mechanism activation than mediation.[7] In some applications, one set of assumptions may be more plausible or defensible than the other, but we cannot make a general claim about the strength of these distinct sets of assumptions. We provide a broader discussion comparing the use of HTEs to mediation analysis in Appendix Appendix B.

4.1 HTEs and Mechanisms with Directly-Affected Outcomes

Collectively, the concept of MIVs and the two exclusion assumptions convey the basic intuition that our initial results draw upon. For HTEs to provide information about a mechanism, it must be the case that the moderator of interest affects: (1) the degree to which treatment activates a mechanism or (2) the mechanism's effect on our outcome of interest, Y. Recognition of the former scenario—consistent with our stylized example of voter updating—is well-known. The latter opens new possibilities for identifying moderators. The exclusion assumptions suggest that if a covariate, X_k, is a MIV for multiple mechanisms (or a direct effect), it cannot be used to confirm the presence of a given mechanism. This logic is straightforward, but it remains implicit in most uses of heterogeneous treatment effects to detect mechanisms. These intuitions give rise to Proposition 1.

Proposition 1. *Suppose that Y is directly affected by mechanism j and Assumptions 1 and 2 hold with respect to X_k. If HTEs exist with respect to X_k, then $X_k \in \boldsymbol{X}^{MIV}$ for mechanism j.*

Proposition 1 conveys important implications about our ability to use heterogeneous treatment effects to provide evidence for a mechanism. Recall that if X_k is a MIV for mechanism j, $AIE_j(X_k = x) \neq AIE_j(X_k = x')$ for some $x, x' \in X_k$. This provides evidence that mechanism j is active for at least one unit. The exclusion assumptions rule out the possibility that

[7] Mediation analysis attempts to estimate the influence of different mechanisms by decomposing total effects into other causal estimands. To the extent that we care about activation of a mechanism (as in HTEs analysis), we can say that a mechanism is activated in mediation analysis if its indirect effect is distinguishable from zero.

$AIE_{\neg j}(X_k = x) \neq AIE_{\neg j}(X_k = x')$ for all other mechanisms ($\neg j$), as well as the possibility that $ADE(X_k = x) \neq ADE(X_k = x')$. When these assumptions hold, HTEs in X_k are thus sufficient to show that X_k is a MIV for mechanism j. This conforms to standard interpretations that the presence of HTEs support arguments about mechanism activation. Nevertheless, this interpretation invokes the two exclusion assumptions, which are generally not invoked explicitly. We now consider the converse: the case when there exist no HTEs with respect to X_k.

Proposition 2. *Suppose that Y is directly affected by mechanism j and Assumptions 1 and 2 hold. If no HTEs exist with respect to X_k, at least one of the following must be true:*

1. *$X_k \notin \boldsymbol{X}^{MIV}$ for mechanism j.*

2. *No MIV exists.*

Proposition 2 shows that a lack of HTEs provides less information with regard to mechanism activation than is generally asserted. Under the exclusion assumptions, there are two reasons why HTEs may not exist with respect to a covariate, X_k. First, it may be the case that X_k is not a MIV for mechanism j. In this sense, we have misspecified the theoretical relationship between a given covariate and a mechanism. Second, it may be the case that no MIV exists for mechanism j. As we discuss in Corollary 1, there are two possible reasons why a MIV would not exist for mechanism j. Importantly, we show that this could happen with an active or an inert mechanism j.

Corollary 1. *If no MIV exists for a mechanism j, there are two possibilities:*

(1) Mechanism j is not active.

(2) Mechanism j is active, but there exists no X for which $AIE_j(X = x) \neq AIE_j(X = x')$.

Case (1) of Corollary 1 is implied by the definition of MIV. If a mechanism is inert—thereby producing an indirect effect of zero for all units—there cannot exist any MIVs, measured or unmeasured. In contrast, in Case (2), a mechanism can be active and produce the same indirect effect for all units. In this case, there are no covariates that moderate the indirect effect. These results show that, in contrast to standard interpretation, a *lack* of heterogeneity cannot tell us about

whether a mechanism is active. It could be active or inactive. Moreover, our theory could be mis-specified, meaning that our postulated MIV, X_k is not actually a MIV. An assessment of HTEs with respect to a single moderator cannot distinguish between these three possibilities. Nor can we assign probabilities to the these (non-mutually exclusive) explanations.

Comparing Propositions 1 and 2, under the two exclusion assumptions, the presence of HTEs provides more information with regard to mechanisms than does the absence of HTE. In this sense, relying upon the a lack of heterogeneity to "rule out" a potential mechanism requires much stronger theoretical assumptions than is generally acknowledged. Specifically, we would need to assume that $X_k \in \mathbf{X}^{MIV}$ in order to rule out the activation of mechanism j.[8] Indeed, such an assumption is precisely what we are trying to *learn* from the presence of HTEs in Proposition 1.

5 Indirectly-Affected Outcomes

When outcomes are directly affected by one or more mechanisms under a theoretical model, the existance of HTEs provides evidence that a mechanism is active, as we have shown in Proposition 1. Yet, the situation is more complicated in the case of indirectly-affected outcomes. As in our discussion of the exogenous shocks and voting, we observed heterogeneous treatment effects in both prior beliefs (the mechanism in the model) and in valence (not the mechanism in the model). In this section, we show that this finding is general to situations in which indirectly-affected outcomes are generated by a non-affine transformation of the relevant directly-affected outcome, as noted in Definition 3.

This represents a large class of mappings that are used in most empirical applications. It includes discrete choices made on the basis of comparisons in utility as in the motivating example.

[8]Note that if we assume that $X_k \in \mathbf{X}^{MIV}$ we have also implicitly assumed that there exists an X for which $AIE_j(X = x) \neq AIE_j(X = x')$.

It also applies to mappings from an attitude to a Q-item Likert scale of the form:

$$h(Y) = \begin{cases} 1 & Y \in (-\infty, c_1] \\ 2 & Y \in (c_1, c_2] \\ \vdots & \\ Q & Y \in (c_{Q+1}, \infty), \end{cases} \tag{15}$$

in which c_t are increasing thresholds in the latent attitude. Our focus for this section is outcomes that are indirectly affected such that they are produced by a non-affine mapping from a directly affected outcome. Note that the proofs to Propositions 1 and 2 are written more generally such that they accommodate any non-zero affine transformation.[9]

To understand what HTEs reveal with respect to an indirectly-affected outcome, it is useful to introduce one final concept. We will denote $\mathbf{X}^R$ as the set of pre-treatment covariates with non-zero effects on the directly-affected outcome, Y.[10] Covariates in $\mathbf{X}^R$ can be thought of as "relevant" for predicting outcome Y. It is clear that for any outcome, Y, and mechanism, j, $\mathbf{X}^{MIV} \subseteq \mathbf{X}^R \subseteq \mathbf{X}$. Typically, these subsets will be proper.

We now return to our main question of interest: what do HTEs reveal with regard to mechanisms? Proposition 3 considers the case when there are HTEs in a covariate X_k. Here, we can learn that $X_k \in \mathbf{X}^R$, but this is not informative about whether $X \in \mathbf{X}^{MIV}$, since $\mathbf{X}^{MIV} \subseteq \mathbf{X}^R$. In order to make an inference about mechanism j (under the exclusion assumptions) we need to know whether $X_k \in \mathbf{X}^{MIV}$. If $X_k \in \mathbf{X}^{MIV}$ for mechanism j, then mechanism j is active. If $X_k \notin \mathbf{X}^{MIV}$ mechanism j may or may not be active. The intuition for this result is straightforward. The non-affine mapping from Y to $h(Y)$ "breaks" the additive separability between any $X \in \mathbf{X}^R$ and the indirect effect of mechanism j, IE_j.

Proposition 3. *Suppose that observed outcome $h(Y)$ is a non-affine mapping of directly-affected*

[9] One commonly-invoked linear mapping of outcomes is a Z-score transformation. These transformations are typically used to standardize or index outcome variables.

[10] Formally, if $X \in \mathbf{X}^{rel}$, then there exist $x \in X$ and $x' \neq x \in X$ such that $Y(x) \neq Y(x')$.

outcome Y and Assumptions 1 and 2 hold. If HTEs exist with respect to X_k, then $X_k \in \boldsymbol{X}^R$.

It is useful to consider a simple numerical example of Proposition 3 where $X_k \in \mathbf{X}^R - \mathbf{X}^{MIV}$, meaning that $X_k \in \mathbf{X}^R$ but $X_k \notin \mathbf{X}^{MIV}$. Suppose that we are interested in how a mobilization treatment, $Z_i \in \{0, 1\}$, affects citizens' decision to vote. Consider two covariates that predict turnout: $X_1 \sim \mathcal{N}(0, 1)$ and $X_2 \in \{0, 1\}$. We will further assume that X_1 is a MIV for mechanism 1, such that:

$$M_1(Z, \mathbf{X}) = (1 + Z)X_1$$

Potential voters' utility from voting is given by:

$$U(Z, X) = M_1(Z, X_1) + X_2 = (1 + Z)X_1 + X_2$$

Our observed behavioral outcome—turnout—is a non-affine function of voter utility as follows:

$$h(U(Z, X)) = \begin{cases} 1 & \text{if } (1 + Z)X_1 + X_2 \geq 0 \\ 0 & \text{else} \end{cases}$$

In this case, there is only one mechanism so Assumptions 1-2 hold by construction. Now, suppose that a researcher mistakenly thought that X_2 was a MIV for a mechanism (either M_1 or a non-existent mechanism). Since we have constructed the data generating process, we know that it is not: $X_2 \in \mathbf{X}^R - \mathbf{X}^{MIV}$. Evaluating the CATE of Z on turnout when $X_2 = 1$, we have:

$$\begin{aligned} CATE(X_2 = 1) &= E[h(U(Z = 1, X)) - h(U(Z = 0, X))|X_2 = 1] \\ &= \Pr(2X_1 + 1 > 0) - \Pr(X_1 + 1 > 0) \\ &= \Phi(-1) - \Phi(-\frac{1}{2}) \\ &\approx -0.15 \end{aligned}$$

Note that $\Phi(\cdot)$ is the cdf of the standard normal distribution, which we invoke because $X_1 \sim \mathcal{N}(0,1)$. Using the same approach it is straightforward to see that $CATE(X_2 = 0) = \Phi(0) - \Phi(0) = 0$. So it is clear that we have HTEs in X_2 because $CATE(X_2 = 1) \neq CATE(X_2 = 0)$. Remember that, by construction, X_2 is *not* a MIV, though it is relevant. This numerical example is analogous to the issue that arises with valence in the analysis of our motivating example.

While it is straightforward to construct such many such examples mathematically, it is reasonable to ask, is this plausible in "real world" examples? We cannot answer this question empirically without observing which mechanisms are at work. Indeed, this is the whole problem that quantitative analysis of mechanisms seeks to (indirectly) answer! However, we can conduct simulations that use real data. In Appendix Appendix E, we use 2020 ANES data in a Monte Carlo simulation motivated by the theoretical model of Little, Schnakenberg, and Turner (2022). This paper proposes two mechanisms that account for how citizens update beliefs in response to new information: accuracy and directional motives. Their model implies that partisanship (or ideology) should be a MIV for directional but not accuracy models, which would allow researchers to examine heterogeneity in partisanship to assess the presence of directional motives as a mechanism.

Following this logic, we simulate different treatment effects on a measure of latent attitudes about greenhouse gas regulation. Some simulations allow for directional motives, others shut down this channel. We show that when treatment has a non-zero effect on the latent attitudes for *any* subset of partisans, as we would expect from accuracy motives alone, there are HTEs on a binary measure of preferences for greenhouse gas regulation. This means that we cannot use HTEs to assess the presence of directional motives with this indirectly measured outcome. Further, using the ANES sample, for some simulated effect sizes, there is greater statistical power to detect heterogeneity in partisanship when effects on the latent variable are homogeneous than when they are heterogeneous.

We now return to a final case of our theoretical analysis by asking when we are examining an outcome that may be indirectly affected by mechanism j, what can we learn from a *lack* of HTE? Proposition 4 indicates that in this case, we can infer that $X_k \in \mathbf{X}$. This is obviously a vacuous

result. We already know that $X_k \in \mathbf{X}$ by definition of $\mathbf{X}$. We purposely state a vacuous result to emphasize how little can be ascertained about mechanisms from the non-existence of HTEs when outcomes are indirectly affected by a mechanism.

Proposition 4. *Suppose that observed outcome $h(Y)$ is a non-affine mapping of directly-affected outcome Y and Assumptions 1 and 2 hold. If HTEs do not exist with respect to X_k, then $X_k \in \boldsymbol{X}$.*

Often we make assumptions about the mapping h. For example, the mapping in (15) imposes assumptions about how latent attitudes translate into Likert-scale responses. When we are willing to make such assumptions, we can refine Proposition 4 slightly. Specifically, in Proposition A1, we show that under Assumptions 1-2, for absolutely continuous directly-affected outcome, Y and the non-affine transformation in (15) (for any $Q \geq 2$ categories), if HTEs do not exist with respect to X_k, then $X_k \notin \mathbf{X}^R$. Because $\mathbf{X}^{MIV} \subseteq \mathbf{X}^R$ we know then that $X_k \notin \mathbf{X}^{MIV}$ if $X_k \notin \mathbf{X}^R$. But as in Proposition 2 and Corollary 1, there are multiple possible explanations: our theory about how X_k relates to mechanism j could be wrong or no MIV exists for mechanism j. These possibilities mean that we cannot make an inference about mechanism (non)-activation from the absence of HTEs with respect to X_k.

In sum, our propositions characterize four cases into which we can classify attempts to ascertain mechanism activation from HTEs analysis, as described in Table 3. On the columns, we stratify by whether the outcome is directly affected by the mechanism or whether it is indirectly affected (via some non-affine transformation of the directly-affected outcome). On the rows, we consider whether there exist HTEs in a covariate of interest, X_k. Our results show that, under exclusion assumptions, this strategy provides information about mechanism activation in one case: when HTEs exist for a directly-affected outcome. In the other cases, HTEs provide incomplete—or no—information about the activation of a mechanism of interest.

6 Implications for Applied Research

Our framework and analysis holds a number of implications for applied research that seeks to study causal mechanisms using HTE. We discuss implications and recommendations in four categories:

	Outcome variable is:	
	Directly affected by mechanism j	Indirectly affected by mechanism j
$\exists$ HTEs wrt X_k:	$X_k \in \mathbf{X}^{MIV}$ $\Longrightarrow M_j$ is active.	$X_k \in \mathbf{X}^{R}$ M_j active or inactive
$\nexists$ HTEs wrt X_k:	$X_k \notin \mathbf{X}^{MIV}$ **and/or** $\nexists$ MIV for mechanism j M_j active or inactive	$X_k \in \mathbf{X}$ (vacuous) M_j active or inactive

Table 3: In this table, $\mathbf{X}^{MIV}$ is defined with respect to mechanism j. The results in each cell invoke Assumptions 1 and 2.

1. Desiderata for applied theory in empirical research

2. Improvements in the interpretation of HTE

3. Recommendations for prospective research design

4. Benefits and limitations of stronger theoretical assumptions

6.1 Three essential theoretical questions

Our framework suggests that some form of theory is necessary to link HTEs with respect to a covariate to a causal mechanism. While our analysis is ultimately agnostic with respect to the *type* of theory (e.g., formal or informal), our framework lays out three attributes of a theory that are needed to support any analysis of causal mechanisms using HTE.

1. *A set of candidate mechanisms*. Researchers must generate a list of the candidate mechanisms that may mediate the effect of Z on Y.

2. *The relationship between a covariate, X_k, and each candidate mechanism*. This requires answering two questions:

   (a) For which mechanism (j), X_k is a candidate MIV?

   (b) Is Assumption 2 plausible for each of the other candidate mechanisms?

3. *Specifying the relationship between the theoretical outcome of interest and measured outcomes*. Which outcome(s) are directly affected by the candidate mechanisms versus indirectly affected by those outcomes?

Question #1 is fairly standard in applied empirical research. Researchers often posit one or more mechanisms of interest in addition to alternative explanations. Questions #2 and #3, instead, are much less standard. When researchers assess HTEs, it is rare to discuss the relationship between the moderator of interest and *other* mechanisms even though these assumptions are required for mechanism attribution, as we show. Explicit justification of a moderator of interest as a candidate MIV of a mechanism may facilitate the search for MIVs. As we show in Figure 2, a MIV can moderate the effect of treatment on the mediator *or* the effect of the mediator on the outcome. Since the latter possibility is generally ignored in applied research, these considerations may broaden the set of candidate MIVs.

Question #3 is not a standard consideration in applied research. We typically do not distinguish between directly-affected outcomes (Y's) and indirectly-affected outcomes ($h(Y)$'s). Yet, as we have shown, this distinction is critical to the use of HTEs to provide information about causal mechanisms. In general, formal theoretic treatments permit straightforward evaluation of whether $h(\cdot)$ is a non-zero linear function, which can tell us whether HTEs could be informative about mechanisms. More broadly, however, Question #3 shows that we may want to evaluate HTEs for some outcomes but not others and provides a principled justification for this determination.

6.2 Improving the interpretation of HTEs as mechanism tests

This framework provides guidance for the *interpretation* of HTEs when researchers are trying to make inferences which causal mechanisms are active. First, while the presence of HTEs provide evidence that a mechanism is active when the exclusion assumptions hold and an outcome is directly affected, the *absence* of HTEs is less informative (even under the same exclusion assumptions). In this sense, a lack of HTEs cannot be used to "rule out" a candidate mechanism or show that it is inert. As Proposition 2 shows, even when an outcome is directly-affected by a mechanism, a lack of HTEs could mean that (1) our model of the relationship between X_k and mechanism j is wrong; or (2) that mechanism j is inert. Because we cannot rule explanation (1), we cannot affirm explanation (2).

This observation is particularly stark when we consider the *statistical* properties of HTE. Low

statistical power for interactions reduces our ability to statistically detect HTEs that do exist. In other words, we risk many false-negatives in inferences related to the existence of heterogeneity. Because a *lack* of HTEs is uninformative about mechanisms, low power suggests that applied researchers often operate in a world in which heterogeneity analysis is unlikely to provide information to support inferences about mechanisms.[11]

6.3 Guidance for research design

Our framework posits several recommendations for the design of causal research that seeks to test mechanisms quantitatively using HTE. Our suggestions are premised on improvements in measurement. In terms of covariates, we are primarily concerned with which covariates are measured and the number of candidate MIVs (per mechanism) among those covariates. Covariates are only useful for ascertaining mechanisms when (1) they are plausibly MIVs for a single mechanism; and (2) they do not moderate direct effects. This observation suggests that special care must be taken when positing candidate MIVs. When pre-treatment covariates are (largely) collected in baseline data collection, there is a need to posit MIVs and defend exclusion assumptions *ex-ante*. Such considerations require more theory and justification than are typically conveyed in the specification of moderation analyses in pre-analysis plans.

When considering directly-affected outcomes, it is very useful to have multiple candidate MIVs for a given mechanism. To see why, consider the case in which we have two candidate MIVs, X_k and X_{k+1} for mechanism j, and both exclusion assumptions hold for both candidate MIVs. Suppose that there do not exist HTEs in X_k but there do exist HTEs in X_{k+1}. If we only measured HTEs with respect to X_k, following Proposition 2, we would not be able to ascertain whether the problem is with the theory ($X_k \notin \mathbf{X}^{MIV}$) or whether there simply exist no MIV for mechanism j. If there exist HTEs in X_{k+1}, we can eliminate the possibility that there do not exist MIV for mechanism j. This would suggest that the theory with respect to X_k is misspecified. This is

[11] Selective reporting of significant results (due to p-hacking or publication bias) complicates the situation further. In this case, evidence in favor of treatment-effect heterogeneity is more likely to be a false-positive, which increases the the probability that researchers infer that a mechanism is active when it is not.

useful insofar as it allows us to make an inference that mechanism j is active. Note, however, that in order to leverage multiple candidate MIVs, both exclusion assumptions must hold for each candidate MIV, which can be quite demanding.

Our distinction between directly- and indirectly-affected outcomes suggests two recommendations for research design. First, if a goal of a research design is to distinguish causal mechanisms, directly-affected outcomes should be prioritized in HTEs analysis. This requires researchers to make clear which outcomes are directly-affected and emphasizes the value of measuring these outcomes. In our motivating example, researchers would typically measure (self-reported) vote choice in a survey to measure the effects of a shock on incumbent support. But if they were running a survey, they could, in principle, elicit willingness-to-pay for the incumbent to try to directly measure voter uility. Our results suggest that the latter would be a worthwhile—if non-standard—investment because HTEs can provide some information about mechanisms with this latter outcome (but not the former).

Second, these results merit broader consideration of latent variable measurement models (Fariss, Kenwick, and Reuning, 2020). It is rare for researchers to explicitly measure treatment effects on estimates of a latent variable (e.g., attitudes or preferences). However, various methods for indexing multiple outcome measures, including Z-score indices (i.e., Kling, Liebman, and Katz, 2007), arguably do this implicitly. More explicit consideration of which latent variables are directly affected by (a) mechanism(s) and how will improve the use of HTEs for mechanism detection. Such considerations can also provide information about which latent variable models are most appropriate.

6.4 Strengthening assumptions provides limited traction on mechanism tests using HTE

Our results suggest that HTEs or lack thereof are uninformative about mechanism activation when measured outcomes are indirectly affected by treatment. It is worthwhile to consider whether we can make progress in this common case by imposing stronger assumptions on CATEs and HTEs. One assumption that is widely utilized in partial identification results is that of monotonicity (Manski, 1997). In our context, monotonicity holds that for all $x' > x \in X_k$, $CATE(x') \geq (\leq$

$)CATE(x)$.

In Appendix D, we show that the assumption of montonicity is not, in general, sufficient to provide information about mechanism activation through analysis of HTE. The mapping between an assumption of montonicity and its implications for HTEs depends on the data generating process that generates directly- and indirectly-affected outcomes. We show this in the context of the following oft-used data generating process for binary outcomes:

$$Y = g(Z, X_k) + \varepsilon_i$$

$$h(Y) = \begin{cases} 0 & \text{if } Y \leq c \\ 1 & \text{if } Y > c, \end{cases}$$

for some constant $c \in (-\infty, \infty)$ and random variable ε distributed according to the density function $f_\varepsilon(\cdot)$. Under the assumption that g is continuous and differentiable, this data-generating process holds that X_k is a MIV if $\frac{\partial^2 g}{\partial Z \partial X_k} \neq 0$ for some $x \in X_k$. An assumption of monotonicity strengthens this condition by further implying that $\frac{\partial^2 g}{\partial Z \partial X_k}$ is weakly positive or negative for all $x \in X_k$. Our goal is to learn whether X_k is a MIV by assessing whether $\frac{\partial^2 g}{\partial Z \partial X_k}$ is non-zero through estimation of HTE.

In Proposition A2, we show that montonicity alone is not sufficient to ensure that HTEs take different signs when $\frac{\partial^2 g}{\partial Z \partial X_k} = 0$ when X_k is not a MIV, and $\frac{\partial^2 Y}{\partial Z \partial X_k} > (<) 0$ when X_k is a MIV and montonicity holds. These results show that we need additional assumptions on the distribution f_ε and/or the functional form of $g(Z, X_k)$ for monotonicity to provide sufficient information to distinguish mechanisms.

Of course, these results consider only one data generating process when considering the implications of assuming monotonicity. But a single example shows, in general, this class of assumptions cannot, in isolation, allow us to distinguish mechanisms through HTE. If researchers are willing to specify a functional form for directly-affected outcome Y and mapping $h(Y)$, it may well be possible to generate a set of sufficient conditions for learning about mechanisms through

HTEs on the outcome $h(Y)$. But this requires far stronger assumptions than monotonicity.

7 Conclusion

Social scientists routinely estimate HTEs with respect to a pre-treatment covariate with the stated intent of mechanism detection. We show that learning about the the activation of a mechanism using HTEs is less straightforward than conveyed by current practice. Specifically, any link between a covariate (moderator) and a mechanism requires exclusion assumptions, such that that covariate does not moderate the indirect effect of other mechanisms or the direct effect of treatment on an outcome. Even when these assumptions hold, we can only use HTEs to affirm the activation of a mechanism when (1) HTEs exist and (2) the outcome is directly affected by a mechanism. Outside this case, HTEs do not provide sufficient information to show that a mechanism is active or inactive. In this sense, HTEs analysis should not be used to "rule out" mechanisms (or show that they are inert).

While mechanism detection is presently the modal use of HTEs in recent work in political science (see Table 1), it is not the only use of HTE. Our results speak to contexts where learning about mechanisms is the central goal. In current practice, HTEs are also increasingly used for extrapolation of treatment effects to different populations/settings and the targeting of treatments. Our results should not be seen as casting doubt on these applications of HTE, because these applications do not rely on questions of how to attribute observed effects to mechanisms (Slough and Tyson, 2023*a*).

Our analysis raises a number of issues for future research to build upon. In particular, we emphasize the need to distinguish between the level at which mechanisms operate—e.g., on utility—and the outcomes we observe. This distinction has underappreciated implications for multiple quantitative methods to detect mechanism activation including mediation analysis and investigation of the sign of treatment effects. We also provide a framework to answer other questions about HTEs and questions of causal structure. This framework can help to clarify the relationship between causal mechanisms and other applications of HTEs to clarify the theoretical foundations of

these approaches.

References

Abramson, Scott F, Korhan Koçak, and Asya Magazinnik. 2022. "What do we learn about voter preferences from conjoint experiments?" *American Journal of Political Science* 66 (4): 1008–1020.

Achen, Christopher H., and Larry M. Bartels. 2017. *Democracy for Realists: Why Elections Do Not Produce Responsive Government*. Princeton, NJ: Princeton University Press.

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

Ashworth, Scott, Christopher R Berry, and Ethan Bueno de Mesquita. 2021. *Theory and Credibility: Integrating Theoretical and Empirical Social Science*. Princeton University Press.

Ashworth, Scott, Christopher R. Berry, and Ethan Bueno de Mesquita. 2023. "Modeling Theories of Women's Underrepresentation in Elections." *American Journal of Political Science* Early View.

Ashworth, Scott, Ethan Bueno de Mesquita, and Amanda Friedenberg. 2018. "Learning about voter rationality." *American Journal of Political Science* 62 (1): 37–54.

Athey, Susan, Julie Tibshirani, and Stefan Wager. 2019. "Generalized random forests." *Annals of Statistics* 47 (2): 1148–1178.

Athey, Susan, and Stefan Wager. 2021. "Policy Learning with Observational Data." *Econometrica* 89 (1): 133–161.

Baccini, Leonardo, Abel Brodeur, and Stephen Weymouth. 2021. "The COVID-19 Pandmeic and the 2020 US Presidential Election." *Journal of Population Economics* 34 (2): 739–767.

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* 114 (2): 375–391.

Coppock, Alexander. 2022. *Persuasion in Parallel*. Chicago, IL: University of Chicago Press.

Devaux, Martin, and Naoki Egami. 2022. "Quantifying Robustness to External Validity Bias." Working paper available at `https://naokiegami.com/paper/external_robust.pdf`.

Dunning, Thad. 2012. *Natural experiments in the social sciences: a design-based approach*. New York: Cambridge University Press.

Egami, Naoki, and Erin Hartman. 2022. "Elements of external validity: Framework, design, and analysis." *American Political Science Review* Forthcoming.

Fariss, Christopher J, Michael R. Kenwick, and Kevin Reuning. 2020. *The SAGE Handbook of Research Methods in Political Science and International Relations*. Number 20 Thousand Oaks, CA: SAGE Publications chapter Measurement Models, pp. 353–370.

Fink, Günther, Margaret McConnell, and Sebastian Vollmer. 2014. "Testing for heterogeneous treatment effects in experimental data: falsediscovery risks and correction procedures." *Journal of Development Effectiveness* 6 (1): 44–57.

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

Grimmer, Justin, Solomon Messing, and Sean J. Westwood. 2017. "Estimating Heterogeneous Treatment Effects and the Effects of Heterogeneous Treatments with Ensemble Methods." *Political Analysis* 25 (4): 413–434.

Healy, Andrew J., Neil Malhotra, and Cecilia Hyunjung Mo. 2010. "Irrelevant events affect voters' evaluations ofgovernment performance." *Proceedings of the National Academy of Sciences* 107 (29): 12804–12809.

Healy, Andrew, and Neil Malhotra. 2010. "Random Events, Economic Losses, and Retrospective Voting: Implications for Democratic Competence." *Quarterly Journal of Political Science* 5 (2): 193–208.

Imai, Kosuke, Luke Keele, and Dustin Tingley. 2010. "A general approach to causal mediation analysis." *Psychological methods* 15 (4): 309.

Imai, Kosuke, Luke Keele, Dustin Tingley, and Teppei Yamamoto. 2011. "Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies." *American Political Science Review* 105 (4): 765–789.

Imai, Kosuke, and Teppei Yamamoto. 2013. "Identification and sensitivity analysis for multiple causal mechanisms: Revisiting evidence from framing experiments." *Political Analysis* 21 (2): 141–171.

Kitagawa, Toru, and Aleksey Tetenov. 2018. "Who Should be Treated? Empirical Welfare Maximization Methods for Treatment Choice." *Econometrica* 86 (2): 591–616.

Kling, Jeffrey R., Jeffrey B. Liebman, and Lawrence R. Katz. 2007. "Experimental Analysis of Neighborhood Effects." *Econometrica* 75 (1): 83–119.

Lee, Soohyung, and Azeem M. Shaikh. 2014. "Multiple Testing and Heterogeneous Treatment Effects: Re-Evaluating the Effect of Progresa on School Enrollment." *Journal of Applied Econometrics* 29: 612–626.

Levy Paluck, Elizabeth. 2010. "The Promising Integration of Qualitative Methods and Field Experiments." *The ANNALS of the American Academy of Political and Social Science* 628 (1): 59–71.

Little, Andrew T., Keith E. Schnakenberg, and Ian R. Turner. 2022. "Motivated Reasoning and Democratic Accountability." *American Political Science Review* 116 (2): 751–767.

Malis, Matt, and Alastair Smith. 2021. "State Visits and Leader Survival." *American Journal of Political Science* 65 (1): 241–256.

Manski, Charles F. 1997. "Monotone Treatment Response." *Econometrica* 65 (6): 1311–1334.

McClelland, Gary H., and Charles M. Judd. 1993. "Statistical Difficulties of Detecting Interactions and Moderator Effects." *Psychological Bulletin* 114 (2): 376–390.

Moscowitz, Daniel. 2021. "Local News, Information, and the Nationalization of U.S. Elections." *American Political Science Review* 115 (1): 114–129.

Neyman, Jerzy. 1923. "Sur les applications de la theorie des probabilites aux experiences agricoles: essai des principes (Masters Thesis); Justification of applications of the calculus of probabilities to the solutions of certain questions in agricultural experimentation. Excerpts English translation (Reprinted)." *Statistical Science* 5: 463–472.

Nilsson, Anton, Carl Bonander, Ulf Strömberg, and Jonas Björk. 2021. "A directed acyclic graph for interactions." *International Journal of Epidemiology* 50 (2): 613–619.

Poole, Keith T., and Howard Rosenthal. 1985. "A Spatial Model for Legislative Roll Call Analysis." *American Journal of Political Science* 29 (2): 357–384.

Rubin, Donald B. 1974. "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of educational Psychology* 66 (5): 688.

Samii, Cyrus. 2016. "Causal empiricism in quantitative research." *The Journal of Politics* 78 (3): 941–955.

Slough, Tara. 2022. "Phantom Counterfactuals." *American Journal of Political Science* forthcoming.

Slough, Tara, and Scott A. Tyson. 2023*a*. "External Validity and Evidence Accumulation." Book manuscript.

Slough, Tara, and Scott A. Tyson. 2023*b*. "External Validity and Meta-Analysis." *American Journal of Political Science* First View.

Slough, Tara, and Scott A. Tyson. 2023*c*. "Sign-Congruent External Validity and Replication." Working paper, available at `http://taraslough.com/assets/pdf/sc_ev_r.pdf`.

Wolfers, Justin. 2002. "Are Voters Rational? Evidence from Gubernatorial Elections." Working paper available at `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=305740`.