American Political Science Review (2026) 1–18
doi:10.1017/S0003055426101580 © The Author(s), 2026. Published by Cambridge University Press on behalf of American Political Science Association. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

Heterogeneous Treatment Effects and Causal Mechanisms

JIAWEI FU *Duke University, United States*
TARA SLOUGH *New York University, United States*

The credibility revolution advances the use of research designs that permit the identification and estimation of causal effects. However, understanding which mechanisms produce measured causal effects remains a challenge. The dominant current approach to the quantitative evaluation of mechanisms relies on the detection of heterogeneous treatment effects (HTEs) with respect to pretreatment covariates. This article develops a framework to understand when the existence of such HTEs can support inferences about the activation of a mechanism. We show first that this design cannot provide evidence of mechanism activation without additional, generally implicit, exclusion assumptions. Further, even when these assumptions are satisfied, the presence of HTEs supports the inference that the mechanism is active but the absence of HTEs is generally uninformative about mechanism activation. We provide novel guidance for interpretation and research design in light of these findings.

Social scientists often make claims about *how* or *why* a treatment affects a given outcome. Researchers might tell us, for example, not only that an outreach program increased COVID-19 reporting rates, but that it did so *by* building trust in government. Or researchers might claim that access to local media increases knowledge of local politics because of the content of news coverage, *not* because of local political advertising. Claims of this type are often made on the basis of heterogeneous treatment effects (HTEs): the effect of the outreach program on COVID-19 reporting is larger in places that initially distrusted government (Haim, Ravanilla, and Sexton 2021), for example, or the effect of local media on knowledge of local politics does not change over the election cycle (Moscowitz 2021). In fact, a majority of recent articles that report HTEs interpret them as evidence about *how* or *why* a treatment affects a given outcome.

But the theoretical foundation for making such claims lags far behind empirical practice. While we have strong theoretical foundations for making claims about treatment effects themselves, theory provides little guidance about the relationship between HTEs and tests of mechanisms. We provide that guidance, developing a theoretical framework that clarifies the assumptions that researchers need to make in order to use HTEs as evidence for or against potential mechanisms. Our analysis reveals that current practice is often misleading—but also that, as is so often the case, making implicit assumptions explicit can correct such errors in future work.

This framework is important because political scientists so often use HTEs to make claims about causal mechanisms. Surveying 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 find that a majority (56%) of quantitative studies estimate HTEs and that, conditional on reporting any HTEs, the vast majority of articles (95%) interpret them as providing information about mechanisms. Taken together, these figures indicate that more than half (53%) of recent quantitative empirical articles in these journals use HTEs to assess mechanisms.[1] Beyond our analysis, Blackwell, Ma, and Opacic (2024) show that HTEs are the modal way that authors in these journals analyze causal mechanisms.

Clearly, the use of HTEs for mechanism attribution is quite common. However, it may be the case that these tests are viewed as secondary in importance to the "main effect" or the treatment effect in the full sample. While there is likely variation in the extent to which readers value main effects versus HTEs, authors commonly emphasize these mechanism tests as important results. The right column of Table 1 shows that conditional on relying on HTEs for mechanism evaluation, a majority (59%) of articles mention these results about mechanisms in the abstract. Given word constraints, we view this as evidence that authors place non-trivial weight on HTE-based tests of mechanisms as important results of their analyses.

We ask an essential but as-yet-unanswered question: under what conditions do HTEs provide evidence of

Jiawei Fu, Assistant Professor, Department of Political Science, Duke University, United States, jiawei.fu@duke.edu.
Corresponding author: Tara Slough, Associate Professor, Department of Politics, New York University, United States, tara.slough@nyu.edu.

Handling editor: Andrew Eggers.

Received: January 10, 2025; revised: June 23, 2025; accepted: February 17, 2026.

[1] Table A1 in the Supplementary Material further documents that the share of studies that use of HTE for mechanism detection is similar across all quantitative research designs/identification strategies in common usage.

https://doi.org/10.1017/S0003055426101580 Published online by Cambridge University Press

TABLE 1. Classification of the Uses of HTEs in All Articles Published in Three Leading Political Science Journals in 2021

	Number of articles:			Prevalence of HTEs		Importance
Journal (Volume)	Total	Quantitative empirical	Reporting HTEs	Pr(Report HTEs | Quant. empirical)	Pr(Mechanism test | Report HTE)	Pr(In abstract | Mech. test)
AJPS (65)	61	45	24	0.53	0.96	0.57
APSR (115)	102	81	48	0.59	0.98	0.55
JoP (83)	142	109	62	0.57	0.92	0.65
Total	305	235	134	0.56	0.95	0.59

mechanism activation? We do so by extending the workhorse causal mediation framework (Imai, Keele, and Tingley 2010). We define a *mechanism* as an underlying process that influences experience in order to produce a (causal) effect when activated (Slough and Tyson 2024). In order to use HTEs to detect mechanisms, empiricists rely on a measured *moderator*, or pretreatment variable, that is thought to predict the degree to which treatment activates a mechanism and/or the degree to which the mechanism affects the outcome of interest. Our framework provides a minimal structure necessary to link moderators to mechanisms to understand what can be learned from the presence or absence of HTEs. Our results characterize the conditions under which HTEs—specifically, a difference in conditional average treatment effects (CATEs) at different levels of a moderator—are sufficient to show that a specific mechanism is active.

Our framework first elucidates the assumptions that are invoked for mechanism attribution. Specifically, if a covariate moderates the effect of a mechanism, it cannot also be a moderator for any other mechanism(s). If it were, there would be no way to determine which mechanism is responsible for the observed heterogeneity. Our first identification result shows that a difference in CATEs is generically equivalent to the difference in the conditional average indirect effect attributable to a mechanism if and only if exclusion assumptions hold. This means that the covariate does not moderate the effects of other mechanisms. Comparing these assumptions to those invoked by other methods for the quantitative study of mechanisms, namely, the assumption of sequential ignorability in causal mediation analysis, the exclusion assumptions are neither (logically) stronger nor weaker than sequential ignorability. This means that using HTEs for mechanism detection is not more or less agnostic than mediation analysis. However, one crucial benefit of HTEs is that they can provide information about mechanisms without measurement of mediators (Imai et al. 2011).

Learning about a mechanism from HTEs requires more than exclusion assumptions. We introduce the concept of a mechanism detector variable (MDV), a covariate that predicts a stronger activation of a mechanism or a stronger effect of a mechanism on an outcome. Under the exclusion assumptions, the existence of HTEs at different levels of a covariate reveals that the covariate is an MDV. This is, in turn, sufficient to infer that the mechanism in question is active for at least one unit in the sample. This accords with the current interpretation of mechanism tests that rely on HTEs. However, if HTEs do not exist for a covariate that is a candidate MDV, we do not learn whether the mechanism is active or inert. The lack of heterogeneity could be a consequence of misspecification of the theory, a mechanism that produces the same effect for all units, or an inert/inactive mechanism. This limits our ability to rule out the activation of a mechanism using HTEs.

Finally, we show how choices about which outcomes to measure and how to measure them can limit the use of HTEs as a test of mechanisms. Specifically, the exclusion assumptions that we articulate imply that the effects of a mechanism on an outcome should be additively separable from the effects of other mechanisms on that outcome. But common non-affine transformations of the outcome variable—for example, binning, logging, or winsorization—violate this property of additive separability, generating heterogeneity that is not informative about mechanism activation.[2] This finding suggests that the way that we measure the effects of a mechanism can limit our ability to attribute those very effects to the mechanism. It reveals the need to be more explicit about the relationship between mechanisms and measured outcomes than is common practice.

This article makes three principal contributions. First, we introduce a new framework to understand the theoretical relationship between causal mechanisms and treatment effect heterogeneity. A large methodological literature provides guidance on the estimation of HTEs (or "interaction effects") (e.g., Athey, Tibshirani, and Wager 2019; Berry, DeMeritt, and Esarey 2009; Brambor, Clark, and Golder 2006; Grimmer, Messing, and Westwood 2017; Hainmueller, Mummolo, and Xu 2018). More recent contributions employ HTEs for extrapolation, prediction, or targeting of treatments (e.g., Athey and Wager 2021; Devaux and Egami 2022; Egami and Hartman 2023; Huang 2024; Kitagawa and Tetenov 2018). Yet, because these contributions are primarily statistical, they do not facilitate theoretical analysis of the relationship between mechanisms and the (measured) treatment effects that they produce. Our

[2] Our focus on additive separability stems from the widespread current practice of comparing conditional treatment effects across subgroups. Different measures of mechanistic influence will generally imply different functional form assumptions.

https://doi.org/10.1017/S0003055426101580 Published online by Cambridge University Press

primary results on the theoretical properties of HTEs—and the problems with existing practice—therefore complement our understanding important statistical issues associated with the estimation of HTEs, namely, limited statistical power (e.g., McClelland and Judd 1993) and multiple-comparisons problems (Fink, McConnell, and Vollmer 2014; Gerber and Green 2012; Lee and Shaikh 2014).

Second, we expand a growing literature on the theoretical implications of empirical models (TIEMs) (Abramson, Koçak, and Magazinnik 2022; Ashworth, Berry, and Bueno de Mesquita 2021; 2024; Bueno de Mesquita and Tyson 2020; Slough 2023). 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 (formal) theoretical models. This analysis complements recent work by Blackwell, Ma, and Opacic (2024), who make explicit the strong assumptions that underpin the use of treatment effects on intermediate outcomes to discern causal mechanisms. Second, we introduce questions about how measured outcomes relate to theoretical constructs. While measurement is central to recent TIEM work on evidence accumulation in a cross-study environment (Slough and Tyson 2023; 2024; 2025), 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. We illustrate this guidance concretely by analyzing four recent empirical studies about how partisan affinity (or bias) condition voter responses to corruption revelation (Anduiza, Gallego, and Muñoz 2013; Arias et al. 2019; de Figueiredo, Hidalgo, and Kasahara 2023; Eggers 2014). 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 not discussed in applied work. Second, we show how the 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 consider the merits of invoking additional assumptions or statistical models to address some of the issues we identify. Collectively, these suggestions allow practitioners to accurately use—or, when indicated, avoid—HTEs as a quantitative test of mechanisms.

MOTIVATING EXAMPLE: CORRUPTION REVELATION AND VOTING

Despite widespread unpopularity of corruption by public officials in public opinion polls, voters in many democracies routinely re-elect politicians who engage in corruption. One explanation for the prevalence of public corruption in many democracies is that corruption by specific politicians is not observed by voters. If voters were to receive information that an incumbent or candidate were corrupt, then they should exhibit less support for the candidate. Scholars have evaluated the empirical merits of this argument using field experiments (e.g., Dunning et al. 2019), randomized corruption audits (e.g., Ferraz and Finan 2008), survey experiments (e.g., Anduiza, Gallego, and Muñoz 2013), and observational studies (e.g., Eggers 2014). Existing meta-studies document heterogeneity in voter responses to such information (Incerti 2020; Slough 2024). Here, we examine whether such heterogeneity is informative about a voter (dis)taste for corruption mechanism.

Specifically, consider a randomized experiment in which the treatment reveals that an incumbent in a constituency has engaged in corruption. We will say that $c_i = 1$ if voter i receives the corruption information treatment and $c_i = 0$ if they do not receive the information. Voters value multiple attributes of politicians. First, they dislike corruption, albeit to different degrees, where $\lambda_i \in (0, 1)$ captures variation in corruption aversion across voters. Second, voters value partisan alignment with politicians. Specifically, voter i may be aligned $(a_i = 1)$, independent/neutral $(a_i = 0)$, or unaligned $(a_i = -1)$ with an incumbent politician. Finally, voters idiosyncratically value valence characteristics of a politician (e.g., their personality or personal background), which we represent with the random variable ε_i. We assume that ε_i is independent of λ_i and a_i and follows a standard normal distribution. Each voter's utility from a vote for the incumbent is given by

$$u_i = -\lambda_i c_i + a_i + \varepsilon_i.$$

Corruption information enters voters' assessment of the incumbent through their distaste for corruption, λ_i. One can think of the product $\lambda_i c_i$ as akin to a mediator that captures the degree to which this distaste is activated by the information treatment c_i. Ultimately, voters choose between the incumbent and a challenger. Without loss of generality, we normalize a voter's utility from a vote for the challenger to be 0. Therefore, voters will vote for the incumbent if $u_i \geq 0$.

From Theory to Empirical Research Design

Mapping this model onto the empirical research design, we consider two measures of voting outcomes. The first outcome, y_{i1}, measures voters' (expected) utility from the incumbent. It is obviously difficult and rare to measure utility from a candidate directly. However, one could, in principle, ask a voter to evaluate their incumbent on a 0–100 scale (or elicit willingness-to-pay for the incumbent's reelection). As in the model, the voter's utility from a vote for the incumbent is given by

$$y_{i1}(c) = -\lambda_i c_i + a_i + \varepsilon_i. \quad (1)$$

The second outcome, y_{i2}, measures each voter's (self-reported) vote choice for the incumbent. Vote choice is

https://doi.org/10.1017/S0003055426101580 Published online by Cambridge University Press

a commonly measured outcome in the literature on voter behavior. This outcome is given by

$$y_{i2}(c) = \begin{cases} 1 & \text{if } -\lambda_i c_i + a_i + \varepsilon_i \geq 0 \\ 0 & \text{else.} \end{cases} \quad (2)$$

The data-generating process of the model is depicted in Figure A1 in the Supplementary Material. As we report in Table 1, empirical researchers often turn to the estimation of HTEs with respect to a pretreatment covariate to assess the activation of a mechanism. In the context of an experiment, heterogeneity is assessed through the estimation of CATEs at different levels of a pretreatment covariate, X_k. Given outcomes $j \in \{1,2\}$:

$$CATE(y_j, X_k = x) = E[y_{ij}(c_i = 1) - y_{ij}(c_i = 0) | X_{ik} = x]. \quad (3)$$

We will say that treatment effects are heterogeneous if for some $x, x' \in X_k$, where $x \neq x'$, $CATE(y_j, X_k = x) - CATE(y_j, X_k = x') \neq 0$. Importantly, whether treatment effects are *homogeneous* or *heterogeneous* in a given covariate, is fundamentally a qualitative classification. This means that even though researchers are using a quantitative metric to evaluate mechanisms, they are fundamentally making a *qualitative* inference about mechanism activation.

Recall that standard interpretations in the empirical literature use the presence of HTEs as evidence that a mechanism is active and the absence of HTEs to assert that a mechanism is inert. Given our model of voter behavior, we ask: do HTEs provide evidence that the relevant mechanism behind the observed reduction in utility and/or incumbent vote share is indeed voter distaste for corruption? To develop intuitions, we evaluate four HTE combinations using moderators $X_1 = \Lambda$ and $X_2 = A$, where Λ is the set of all possible values of λ_i (corruption aversion) and A is the set of all possible values of a_i (partisan alignment with the incumbent), and outcomes $y \in \{y_1, y_2\}$. Remark 1 shows that for the outcome measuring a voter's utility from a vote for the incumbent (y_1), heterogeneity in CATEs correctly provides evidence that the mechanism is voter distaste for corruption, not some type of corruption-induced partisan-realignment or biased learning that depends on partisanship (e.g., motivated reasoning in Little, Schnakenberg, and Turner 2022) (for example) that are not present in our model.

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

(a) Given $\lambda > \lambda' \in \Lambda$, $|CATE(y_1, X_1 = \lambda)| > |CATE(y_1, X_1 = \lambda')|$.
(b) Given $a \neq a' \in A$, $CATE(y_1, X_2 = a) - CATE(y_1, X_2 = a') = 0$.
(c) If for some $x \neq x'$, $CATE(y_1, X_k = x) - CATE(y_1, X_k = x') \neq 0$, then $X_k = X_1 = \Lambda$.

(All proofs are in the Supplementary Material.)

Researchers will detect heterogeneity with the corruption aversion moderator λ (by a). Voters who are more corruption averse (larger λ) value the incumbent less upon receiving the information than voters with lower corruption aversion. HTEs are not observed for the partisan alignment (non)-moderator a (by b) because ideological alignment (not the mechanism) and distaste for corruption (the mechanism) are additively separable in voters' utility (in Equation 1). Here, c shows that researchers are unlikely to misattribute the mechanism through conventional interpretation 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 partisan alignment, a, as well as for different levels of corruption aversion, λ. A naïve interpretation might suggest that the effect of the corruption information *does* work through some channel involving ideological realignment or biased learning in addition to a channel involving distaste for corruption.

Remark 2. For the outcome measuring voter choice y_2,

(a) For $\lambda > \lambda' \in \Lambda$, $|CATE(y_2, X_1 = \lambda)| > |CATE(y_2, X_1 = \lambda')|$.
(b) $|CATE(y_2, X_2 = -1)| < |CATE(y_2, X_2 = 1)| < |CATE(y_2, X_2 = 0)|$.
(c) If $CATE(y_2, X_k = x) - CATE(y_2, X_k = x') \neq 0$, then $X_k = X_1 = \Lambda$ or $X_k = X_2 = A$.

Voters with stronger aversion to corruption exhibit larger treatment effects on their voting behavior (by a). Interestingly, even though distaste for corruption is the unique mechanism (in the model) through which information affects vote choice, we now observe HTEs across voters of different partisan alignments. Specifically, the treatment effect is largest among neutral/independent voters ($a = 0$), while strongly aligned partisans ($a = -1, 1$) exhibit smaller changes in voting behavior, by b. Moreover, the effect is greater for incumbent-aligned voters than for challenger-aligned voters. This pattern is driven by a measurement concern: ceiling and floor effects that emerge when utility is transformed into vote choice. Aligned voters are substantially more likely to vote for or against the incumbent regardless of new information, while moderates hover around the decision threshold, making them more responsive to changes in observed corruption.

Importantly, the smaller swings observed among strong partisans are not due to the presence of an additional mechanism. Rather, it is because their prior disposition already places them near the upper or lower bounds of the probability scale in terms of propensity to vote for the incumbent. Researchers who are unaware of the underlying mechanism might mistakenly conclude that partisan voters are "less responsive" to treatment and, as a result, wrongly infer that they are engaged in biased learning/insensitive to information or that they are less prone to partisan realignment on the basis of corruption information.

https://doi.org/10.1017/S0003055426101580 Published online by Cambridge University Press

More concretely, we see HTEs in partisan affiliation (a) for vote choice because voters make a binary choice between the incumbent and challenger. But this binary choice means that the distaste for corruption mechanism and the partisan alignment predictor (which does not moderate the mechanism) are no longer additively separable with respect to vote choice. Consequently, researchers are apt to detect HTEs in partisan alignment even when it does not moderate any mechanism. A naïve interpretation of this test would lead to a Type-I error in our inference about the activation of a mechanism involving partisan alignment.

While Remarks 1 and 2 rely on theoretical analysis of a model, one might ask whether an empirical researcher would be likely to detect this heterogeneity in their data. We therefore conduct a Monte Carlo simulation in Appendix A3.3 of the Supplementary Material that uses our theoretical model to guide the data generation process from a hypothetical experiment. We vary: (1) (average) voter support for the incumbent in the electorate ($\bar{y}_{i2} \in \{0.1, 0.2, \ldots, 0.9\}$) by varying the mean of the valence distribution and (2) sample size in the hypothetical experiment ($n \in \{100, 500, 1,000\}$). Consistent with our theoretical results, we observe HTEs in the corruption aversion covariate for both outcomes (voter utility and vote choice) and HTEs in partisan affiliation only for the vote choice outcome. More importantly, with the finite sample sizes that we simulate, we show that for high and low values of voter support for the incumbent in the electorate, we are *more likely (better powered) to detect HTEs in partisan affiliation* than in corruption aversion for the vote choice outcome. This finding holds across multiple estimators of the difference in CATEs (Figures A3 and A5 in the Supplementary Material). Thus, we cannot solely rely on the statistical properties of our designs to help us screen heterogeneity that is informative about mechanism activation from heterogeneity that is not.

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 mechanisms 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 shows that the use of HTEs for assessing mechanism activation depends on the measurement of outcomes of interest.

FRAMEWORK

Defining HTEs

Our framework is built upon the potential outcomes framework or Neyman–Rubin causal model (Neyman 1923; Rubin 1974). We denote a randomly assigned treatment by $Z \in \{z, z'\}$.[3] In order to consider HTEs with respect to pretreatment moderators, denote the vector of pretreatment covariates by $X = (X_1, X_2, \ldots, X_K) \in \mathbb{R}^K$. Some of X_k are *measured*. To improve readability and emphasize the key elements of the framework, we omit subscript i from the notation throughout the framework and results sections.

A valid mediator, or mechanism representation, should: (1) be affected by treatment, Z, and (2) have a nonzero effect on the outcome. Further, the effect of treatment on a mediator or the mediator's effect on the outcome could vary with some covariate(s), X_k. We define a mediator as a function denoted by $M(Z;X)$.[4] It represents the potential outcomes of causal mediator given treatment Z and covariates X. While we employ mediators to clarify the link between HTEs and mediation, these mediators (or mechanisms) do not need to be measured in order to analyze HTEs. Finally, we denote a potential outcome by $Y(Z, M; X)$.

We consider the practice of quantifying HTEs by estimating and comparing CATEs, as documented in Table 1. Following this convention, we consider HTEs with respect to pretreatment moderators, that is, for some variable X_k, where $k \in \{1, 2, \ldots, K\}$.

Definition 1 (Conditional Average Treatment Effect). Consider pretreatment covariate X_k. Given that $z \neq z' \in Z$, the CATE of Z on Y when $X_k = x$ is

$$CATE^Y(X_k = x) = E_{X_{\neg k}}[Y(Z = z, M; X) - Y(Z = z', M; X) | X_k = x].$$

Note that the CATE of a treatment Z is defined with respect to (potential) outcome variable Y. We will index estimands by the outcome variable of interest throughout. Given this definition, there exist HTEs when CATEs of Z on Y differ at different values of a covariate X_k. With some abuse of notation, we will also use X_k to note the support of the covariate.

Definition 2 (Heterogeneous Treatment Effects). HTEs on Y exist with respect to pretreatment covariate X_k if $CATE^Y(X_k = x) \neq CATE^Y(X_k = x')$ for some $x \neq x' \in X_k$.[5]

Definitions 1 and 2 formalize the current practice of comparing CATEs to evaluate whether treatment effects exhibit heterogeneity. However, while our potential outcomes implicitly depend on a mediator, M, there is not yet a link to an underlying mechanism.

[3] All results hold for observational studies in which treatment is conditionally independent of potential outcomes, given a conditioning set of covariates, X.

[4] Note that potential outcomes are often written as a function of only the manipulated treatment, for example, $M(Z)$ and $Y(Z)$, to reflect "no causation without manipulation" (Holland 1986, 959). We choose to denote covariates X as arguments to both potential outcomes in order to clarify the structure of covariates and mediators in our potential outcomes.

[5] More precisely, the probability measure for the set that contains such x and x' is nonzero.

https://doi.org/10.1017/S0003055426101580 Published online by Cambridge University Press

Causal Mediation and Indirect Effects

We now develop the mapping between the analysis of treatment effect heterogeneity and causal mediation, which seeks to quantify the effect of one or more mechanisms. Specifically, consider a decomposition of the total effect (on a unit, i) into direct and indirect effects (Imai and Yamamoto 2013). Suppose that there exist two mediators (or mechanisms), indexed by M_1, M_2. Given two treatment values, $z, z' \in Z$, the total effect of Z on Y is

$$TE^Y(z,z';X) = Y(z,M_1(z;X),M_2(z;X);X) - Y(z',M_1(z';X),M_2(z';X);X). \quad (4)$$

Our notation varies slightly from conventional presentations of mediation that only consider one mechanism (mediator) (Imai, Keele, and Tingley 2010). In the main text, we describe the case with two mechanisms since it is straightforward to generalize this to the special case of one mechanism or to a setting with more than two mechanisms. Further, as above, we continue to index causal effects by the outcome variable, here Y. Treatment effects on Y may consist of direct (DE^Y) and indirect (IE_1^Y and IE_2^Y) effects, as follows:[6]

$$DE^Y(z,z';X) = Y(z,M_1(z;X),M_2(z;X);X) - Y(z',M_1(z;X),M_2(z;X);X) \quad (5)$$

$$IE_1^Y(z,z';X) = Y(z',M_1(z;X),M_2(z;X);X) - Y(z',M_1(z';X),M_2(z;X);X) \quad (6)$$

$$IE_2^Y(z,z';X) = Y(z',M_1(z';X),M_2(z;X);X) - Y(z',M_1(z';X),M_2(z';X);X). \quad (7)$$

The direct effect, $DE^Y(z,z';X)$, represents the direct effect of Z on Y holding both mediators M at potential outcomes $M(z;X)$. Our preferred interpretation of the direct effect is the composite effect of any other mechanisms (aside from M_1 and M_2). However, the direct effect could also include unmediated effects of treatment on an outcome. In our motivating example, there is only one mechanism—voter distaste for observed corruption—so the direct effect (all other mechanisms) is zero. This is evident because the treatment has zero effect if we fix the mediator, $\lambda_i c_i$, to a given level. The indirect effect of mechanism $j \in \{1,2\}$ measures the effect on the outcome that operates by changing the potential outcome of mediator M_j. In our example, the indirect effect measures the effect that passes through the distaste mechanism.

[6] See Acharya, Blackwell, and Sen (2016) for the identification of the controlled direct effect.

As is standard, we can rewrite the total effect as follows:[7]

$$TE^Y(z,z';X) = DE^Y(z,z';X) + IE_1^Y(z,z';X) + IE_2^Y(z,z';X), \quad (8)$$

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

$$ATE^Y(z,z') = E_X\Big[Y(z,M_1(z;X),M_2(z;X);X) - Y(z',M_1(z';X),M_2(z';X);X)\Big] \quad (9)$$

$$= E_X[DE^Y(z,z';X) + IE_1^Y(z,z';X) + IE_2^Y(z,z';X)] \quad (10)$$

$$=: ADE^Y(z,z') + AIE_1^Y(z,z') + AIE_2^Y(z,z'). \quad (11)$$

We use ADE^Y and AIE_j^Y to denote the average direct effect and the average indirect effect of mechanism j for outcome Y, respectively. Throughout the article, we assume that the expectation in Equation 10 is well defined.

HTEs and the Identification of Indirect Effects

We call a mechanism *active* if there exists at least one unit with covariates $X = x$ for whom the indirect effect is nonzero for given treatment values $z, z' \in Z$.

Definition 3 (Active Mechanism). Given two treatment values $z, z' \in Z$, mechanism j, represented by mediator M_j, is active if there exists at least one $x \in X$ such that $IE_j^Y(z,z';X=x) \neq 0$.[8]

How do HTEs—differences in CATEs—relate to the indirect effects that are estimated within the mediation framework? To show this relationship, consider a decomposition of a CATE into a conditional ADE and two conditional AIEs following Equation 11:[9]

$$CATE^Y(X_k = x) = ADE^Y(z,z';X_k = x) + \sum_{j=1}^{2} AIE_j^Y(z,z';X_k = x).$$

One can then express the difference in CATEs at two distinct levels of X_k as

[7] Here, we assume that the direct and indirect effects do not vary at different levels of Z. See more discussion by Imai et al. (2011). All results in the article hold with other decompositions, albeit with slightly different interpretations. See additional discussions of general cases in Appendix D of the Supplementary Material and correlated mechanisms in Appendix E of the Supplementary Material.

[8] Within the potential outcomes framework, the value of X is fixed for each unit in the population. Therefore, this definition is equivalent to saying that there exists at least one unit with $X = x$ for whom the indirect effect of mechanism j is nonzero.

[9] Conditional $ADE^Y(z,z';X_k = x)$ and $AIE_j^Y(z,z';X_k = x)$ evaluate expectations over all X but fix X_k at x.

https://doi.org/10.1017/S0003055426101580 Published online by Cambridge University Press

$$CATE^Y(X_k = x) - CATE^Y(X_k = x') = [ADE^Y(z,z';X_k = x) - ADE^Y(z,z';X_k = x')] + \sum_{j=1}^{2}[AIE_j^Y(z,z';X_k = x) - AIE_j^Y(z,z';X_k = x')]. \quad (12)$$

From Equation 12, it is clear that HTEs could arise from differences in direct and/or indirect effects. Suppose that we were interested in evaluating the activation of mechanism 1 with respect to outcome Y. Equation 12 shows that we cannot automatically attribute observed heterogeneity to mechanism 1. Instead, to link a difference in CATEs (HTEs) to a difference in indirect effects, we need to assume that the conditional direct effect $ADE^Y(z,z';X_k = x)$ and any conditional indirect effect(s) of the other mechanism(s) $AIE_2^Y(z,z';X_k = x)$ do not vary at different levels of X_k, as stated in Assumptions 1 and 2. Assumption 1 formalizes an exclusion assumption posited by Imai et al. (2011), who focus on the case of a single mechanism.[10]

Assumption 1 (Exclusion I). Given $z,z' \in Z$, X_k is excluded from the direct effect such that $ADE^Y(z,z';X_k = x) = ADE^Y(z,z';X_k = x')$ for all $x,x' \in X_k$.

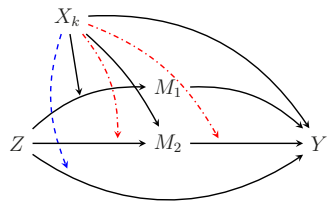
Assumption 2 (Exclusion II). Given $z,z' \in Z$, X_k is excluded from the indirect effect of the other mechanism 2: $AIE_2^Y(z,z';X_k = x) = AIE_2^Y(z,z';X_k = x')$ for all $x,x' \in X_k$.

Assumptions 1 and 2 constrain the relationship between a moderator, X_k, other mechanisms (e.g., M_2), and any direct effect of treatment.[11] Figure 1 illustrates these assumptions graphically. While this figure resembles a directed acyclic graph (DAG), we depart from the conventional presentation of DAGs as vertices (nodes) and edges (arrows) because there are edges that point to other edges (rather than vertices). We make this departure because our assumptions impose greater structure on the possible causal moderation (or lack thereof) than is assumed in traditional DAGs. This departure is not new. Nilsson et al. (2021) note that there is no standardized representation of causal moderation in DAGs, so our graphs are informed by an existing proposal for the representation of these effects by Weinberg (2007). This notation allows us to accurately convey the structure of causal moderation.

[10] Specifically, Imai, Keele, and Tingley (2010, 784) write "if the size of the ADE does not depend on the pretreatment covariate … a statistically significant interaction term implies that the [average causal mediation effect] is larger for one group … than for another group."

[11] Other methods for mechanism detection invoke distinct exclusion assumptions. For example, the scaling stage of implicit mediation uses instrumental variables analysis to estimate causal mediation effects (Bullock and Green 2021). In contrast to our analysis of HTEs, the exclusion assumptions underpinning implicit mediation restrict the mediators that might be activated by a treatment. See also Fu (2024) for a related discussion on mediation analysis with HTEs.

FIGURE 1. Visualization of Exclusion Assumptions



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

Once Assumptions 1 and 2 are invoked, it is straightforward to see that the difference in CATEs reduces to differences in AIE_1^Y at different levels of X_k, which we state formally in Proposition 1. Assumptions 1 and 2 are generically necessary because it is possible that $ADE^Y(x) - ADE^Y(x') \neq 0$ and $AIE_2^Y(x) - AIE_2^Y(x') \neq 0$ exactly offset each other. But under generic parameter values, the probability of this knife-edge event is zero.

Proposition 1. Assumptions 1 and 2 are sufficient and generically necessary for $CATE^Y(X_k = x) - CATE^Y(X_k = x') = AIE_1^Y(z,z';X_k = x) - AIE_1^Y(z,z';X_k = x')$, for all $x,x' \in X_k$.

Proposition 1 clarifies that a difference in CATEs does not identify either conditional AIEs of mechanism 1 ($AIE_1^Y(z,z';X_k = x)$ or $AIE_1^Y(z,z';X_k = x')$) in the absence of further assumptions. Rather, the difference in CATEs identifies a *difference* in conditional AIEs. Thus, identification of this difference is not sufficient to identify indirect effects, as is the goal in (standard) mediation analysis. However, it is straightforward to see that if the difference in AIEs is not equal to zero, there must exist some unit for whom the AIE is not equal to zero. A nonzero difference in AIEs is therefore a sufficient condition for the activation of the relevant mechanism for at least one unit. This identification result motivates a more precise version of our research question: "Under what conditions are HTEs with respect to a covariate X_k sufficient to show that mechanism is active?"

To understand our later results, it is useful to see how Assumptions 1 and 2 could be violated. The first and most obvious violation would be that a covariate X_k moderates multiple mechanisms (or one mechanism and the direct effect). This is clear from Figure 1. A second and less obvious violation is that the additive separability of the (indirect) effects of mechanisms breaks down. If this were to occur, any X_k that moderates the effect of mechanism 1 must also moderate the effect of any other mechanism that "interacts with" or whose (indirect) effect depends on the effect of mechanism 1. In Figure 1, this would occur if, for example, there was an interaction between the effects of M_1 and M_2 on the outcome Y.

It is important to note that mediation analysis does not invoke Assumption 1 or 2, and instead invokes an

https://doi.org/10.1017/S0003055426101580 Published online by Cambridge University Press

assumption of sequential ignorability (Imai, Keele, and Yamamoto 2010). 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 or less agnostic test of mechanism activation than mediation. 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 E of the Supplementary Material.

Connecting Mechanisms to Measured Variables

While our identification results show that HTEs *can* provide information about mechanism activation for some unit(s), the framework does not yet elucidate the relationship between a mechanism and measured variables. To understand why it is critical to develop this relationship beyond our identification results, note that a randomly-generated covariate that is independent of all variables in a research design (or system) satisfies Assumptions 1 and 2 by construction. Here, we would not expect to observe HTEs at different levels of the randomly-generated variable. But this lack of heterogeneity should not be informative about the substantive mechanism(s) at play.

We view a mechanism as an underlying process that responds to some activation and produces a given set of outputs. In the context of our running example, the voter distaste mechanism is activated by the observation of corruption information about the incumbent. It produces a number of outputs—a voter's assessment of the incumbent and their voting decision—among other (unmodeled) possibilities. None of these objects—a voter's information, their utility, or their voting decision—needs be inherently quantitative (though they could be). As social scientists, we choose how to measure and operationalize each of these objects: we normalize the voter information treatment to a binary (0/1) scale for convenient estimation of treatment effects; we choose some type of Likert scale to measure voter assessments/utility; and we choose self-reported vote choice or aggregated voting results to measure vote decisions. These operationalizations facilitate our ability to measure or quantify the effect of a mechanism on an outcome using, for example, a difference in CATEs. This implies that our ability to observe a mechanism's effect depends fundamentally on *how* we choose to measure it (Slough and Tyson 2024). Our concern here is therefore the link between a substantive mechanism and the measured variables in our framework.

Outcome Variables and Mechanisms

First, consider the relationship between measured outcomes and the outputs of a mechanism. A given mechanism produces multiple possible outputs; a researcher chooses to operationalize and measure some subset of those outputs as outcome variables. But not all outcome variables relate to the underlying mechanism in the same way. For example, in our motivating example, the mechanism—voter distaste for observed corruption—affects a voter's utility from the incumbent (y_1). The second outcome, vote choice, is a deterministic but non-affine function of utility given by the function $y_2(c) = \mathbb{I}[y_1(c) \geq 0]$. We refer to $y_2(c)$ as a *transformed (potential) outcome* relative to y_1.

Definition 4. Given a (potential) outcome $Y(Z,M;X)$. Let $\tilde{Y}(Z,M;X) = h(Y)$, where $h(\cdot)$ is a non-affine function. Then we call $\tilde{Y}(Z,M;X)$ a *transformed (potential) outcome* relative to $Y(Z,M;X)$.

Formally, for a function $h(\cdot)$, we define the increment $\delta(t;d) = h(t+d) - h(t)$. We say that the function h is non-affine if there exists nondegenerate t, t', and $d \neq 0$, such that the increment is different, $\delta(t;d) \neq \delta(t';d)$.[12] This is a general definition that does not assume differentiability of the function $h(\cdot)$.

The non-affine transformation is important because it affects the validity of the exclusion assumptions. Specifically, suppose that we believed that the exclusion assumptions held for a given covariate, X_k, mechanism, M_1, and outcome Y.[13] Given our definition of the non-affine function $h(\cdot)$, $ADE^{\tilde{Y}}(z,z';X_k = x) = \mathbb{E}[\delta(Y(z',M_1,M_2;X_k = x); Y(z,M_1,M_2;X_k = x) - Y(z',M_1,M_2;X_k = x))]$. However, in general, $ADE^{\tilde{Y}}(z,z';X_k = x) \neq ADE^{\tilde{Y}}(z,z';X_k = x')$ without other specific assumptions about the functional form of $h(\cdot)$. This means that even if Assumption 1 holds for Y, it generally will not hold for the transformed outcome $\tilde{Y}$. A similar result holds for Assumption 2 and $AIE_2^{\tilde{Y}}$. The logic for these observations is that the non-affine transformation "breaks" the additive separability of the mechanism from (1) other mechanism(s) and (2) predictors of the outcome. This mechanically generates additional causal moderation with different mechanisms without changing the underlying processes through which treatment affects the outcome. Given our distinction between substantive mechanisms and measurement of the effects they produce, we view the introduction of heterogeneity via a change in outcome variables as distinct from the presence of heterogeneity that exists in how mechanisms present in the world.

When are transformed outcomes used in applied work? Three cases are quite common (but non-exhaustive). The first is akin to our running example: a treatment induces a change in information or utility that then affects some discrete choice of strategy. The second holds that a treatment changes an actor's attitude (Y). But since attitudes are latent, survey researchers measure changes in the attitude by employing a Likert scale of the form:

[12] We assume that $h(t+d)$ is well-defined.

[13] Note that there is no guarantee that there exists an outcome for which Assumptions 1 and 2 hold in a given system of mechanisms.

https://doi.org/10.1017/S0003055426101580 Published online by Cambridge University Press

$$h(Y) = \begin{cases} 1 & Y \in (-\infty, c_1] \\ 2 & Y \in (c_1, c_2] \\ \vdots & \\ Q & Y \in (c_{Q-1}, \infty), \end{cases} \quad (13)$$

in which c_t denotes increasing thresholds in a latent attitude. Third, a researcher may employ non-affine transformations of an outcome to demonstrate the robustness of results to measurement choices. Here, they may bin a count variable to capture the extensive margin of some behavior or winsorize or logarithmize a skewed outcome, etc. In any of these cases, even if Assumptions 1 and 2 were to hold for the original outcome (utility, attitudes, or the raw variable), they will not hold for the transformed outcome.

Moderators and Mechanisms

Second, consider the role of a measured covariate, X_k, which we seek to use to detect the activation of focal mechanism M_1. An MDV for M_1 is a covariate that produces different conditional AIEs at different values. To economize notation, we will denote $AIE_1^Y(X_k = x) = AIE_1^Y(z, z'; X_k = x, X_{\neg k})$ as the average indirect effect of mechanism (mediator) 1 when $X_k = x$.

Definition 5. A pretreatment covariate X_k is an MDV for mechanism j with respect to outcome Y if for some $x, x' \in X_k$, $AIE_j^Y(X_k = x) \neq AIE_j^Y(X_k = x')$.

We then denote $\mathrm{X}^{MDV} \subseteq \{X_1, X_2, \ldots, X_L\}, L \leq K$, as the (possibly empty) set of covariates that satisfy Definition 5 for the mechanism j. Clearly, if $X_k \in \mathrm{X}^{MDV}$, then mechanism j is active. Therefore, covariate X_k can serve as an indicator for a mechanism/mediator of interest.[14] Under our definition of MDVs, it could be the case that X_k moderates the effect of the treatment on the mediator. 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. Researchers using HTEs to investigate mechanisms seek to detect evidence of a mechanism using treatment-by-covariate interactions with a proposed MDV. In other words, given a proposed MDV X_k, we want to learn whether $X_k \in \mathrm{X}^{MDV}$.

RESULTS

We consider the conditions under which HTEs (or lack thereof) are informative about the activation of a mechanism, M_1, for some unit(s) in a sample. To do so, we analyze four exhaustive and mutually exclusive cases that vary the (1) existence of HTEs and (2) whether an outcome is transformed by a non-affine function or not. In each case, we will assume that Assumptions 1 and 2 hold for the outcome Y. This follows the identification result in Proposition 1. If we were not to invoke these assumptions, we could not link differences in CATEs (HTEs) to differences in the AIEs of a mechanism.

[14] A slightly stronger version of Definition 5 holds when Y is continuously differentiable with respect to M, Z, and X_k. In this case, Definition 5 can be expressed as $\frac{\partial}{\partial X_k}\left(\frac{\partial Y}{\partial M}\frac{\partial M}{\partial Z}\right) \neq 0$ for some $x \in X_k$.

FIGURE 2. Causal Structure of Two MDVs for Mechanism M



Note: Both panels are consistent with Definition 5.

Case 1: HTEs Exist for Outcome $Y(Z)$

Proposition 2 analyzes the case in which HTEs exist (are observed) and Assumptions 1 and 2 are assumed to hold for the outcome, Y. It shows that HTEs can provide evidence that a covariate is an MDV for some mechanism of interest, M_1. Recall that if X_k is an MDV for mechanism 1, $AIE_1^Y(X_k = x) \neq AIE_1^Y(X_k = x')$ for some $x, x' \in X_k$. This is sufficient to provide evidence that mechanism M_1 is active for at least one unit.

Proposition 2. Suppose Assumptions 1 and 2 hold with respect to X_k for outcome Y. If HTEs exist with respect to X_k, then $X_k \in \mathrm{X}^{MDV}$ for mechanism M_1.

This conforms to standard interpretations that the HTEs provide evidence that a mechanism is active. Nevertheless, this finding relies critically upon the validity of exclusion assumptions, following Proposition 1. If one were to detect heterogeneity in an empirical study in this class, the heterogeneity would indeed provide evidence that the postulated mechanism, M_1, is active for some units.

Case 2: HTEs Do Not Exist for Outcome $Y(Z)$

We now consider the converse: the case when there exist no HTEs with respect to X_k and Assumptions 1 and 2 are assumed to hold for the outcome, Y.

Proposition 3. Suppose Assumptions 1 and 2 hold with respect to X_k for outcome Y. If no HTEs exist with respect to X_k, then:

1. $X_k \notin \mathrm{X}^{MDV}$ for mechanism M_1, if $\mathrm{X}^{MDV} \neq \emptyset$.
2. Otherwise, no MDV exists for mechanism M_1.

Proposition 3 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 an MDV for mechanism M_1. This is a failure of research design in which the

https://doi.org/10.1017/S0003055426101580 Published online by Cambridge University Press

researcher has misspecified the theoretical relationship between a given covariate and a mechanism. Second, it may be the case that no MDV exists for mechanism M_1. This is a feature of the mechanism as it exists in the world (rather than the consequence of a researcher's choices). As we discuss in Corollary 1, there are two possible reasons why an MDV would not exist for mechanism M_1. Importantly, we show that this could happen with an active or an inert mechanism M_1.

Corollary 1. If no MDV exists for a mechanism M_1, there are two possibilities:

(1) Mechanism M_1 is not active.
(2) Mechanism M_1 is active, but produces the same effect for all units, so there exists no X_k for which $AIE_1^Y(X_k = x) \neq AIE_1^Y(X_k = x')$.

Case (1) of Corollary 1 is implied by the definition of MDV. If a mechanism is inert—thereby producing an indirect effect of zero for all units—there cannot exist any MDVs, measured or unmeasured. In contrast, in Case (2), a mechanism can be active but it produces the same indirect effect for all units. Recall that a nonzero difference in average indirect effects at different levels of a covariate X_k is a *sufficient* condition for mechanism activation. However, it is not a *necessary* condition for mechanism activation. These results show that, in contrast to standard interpretation, a *lack* of heterogeneity cannot tell us about whether a mechanism is active. Moreover, our theory could be misspecified, meaning that our postulated MDV, X_k is not actually an MDV. An assessment of HTEs with respect to a single moderator cannot distinguish between these three possibilities. Nor can we assign probabilities to these (not mutually exclusive) explanations without stronger assumptions.

Case 3: HTEs Exist for Transformed Outcome $\tilde{Y}(Z)$

To understand what HTEs reveal with respect to a (non-affinely) transformed outcome $\tilde{Y}(Z)$, it is useful to introduce one final concept. We will denote $\mathrm{X}^R = \{X_1, X_2, \ldots, X_K\}$, as the set of all possible pretreatment covariates with nonzero effects on the outcome, Y.[15] Covariates in X^R can be thought of as "relevant" for predicting outcome Y. It is also useful to let X be the set of all possible pretreatment covariates. It is clear that for any outcome, Y, and mechanism M, $\mathrm{X}^{MDV} \subseteq \mathrm{X}^R \subseteq \mathrm{X}$. Typically, these subsets will be proper.

We now return to our main question of interest: what do HTEs reveal with regard to mechanisms? Proposition 4 considers the case when there are HTEs in a covariate X_k. Here, we can learn that $X_k \in \mathrm{X}^R$, but this is not informative about whether $X \in \mathrm{X}^{MDV}$, since $\mathrm{X}^{MDV} \subseteq \mathrm{X}^R$. Why do we observe heterogeneity in covariates in X^R that are not MDVs? The non-affine transformation $h(\cdot)$ generically means that outcome $\tilde{Y}$ does not satisfy Assumptions 1 and 2, even if Y satisfies both assumptions. In this case, the difference in CATEs no longer identifies a difference in the conditional AIEs of interest! Indeed, the difference in CATEs for the transformed outcome is produced by the (sum of) differences in conditional $ADE^{\tilde{Y}}$'s and conditional $AIE^{\tilde{Y}}$'s for all mechanisms. Violation of the identifying assumptions means that we can no longer attribute HTEs to the mechanism of interest, M_1. This means that (absent stronger assumptions), we cannot learn about the activation of M_1 from HTEs on the transformed outcome.

Proposition 4. Suppose Assumptions 1 and 2 hold with respect to X_k for outcome Y. Let observed outcome $\tilde{Y}$ be a transformed outcome relative to Y. If HTEs exist with respect to X_k for outcome $\tilde{Y}$, then $X_k \in \mathrm{X}^R$.

Case 4: HTEs Do Not Exist for Transformed Outcome $\tilde{Y}(Z)$

We now return to a final case of our theoretical analysis by asking when we are examining a transformed outcome that may be affected by mechanism M_1, what can we learn from a *lack* of HTEs? Proposition 5 indicates that in this case, we can infer that $X_k \in \mathrm{X}$. This is obviously a vacuous result. We already know that $X_k \in \mathrm{X}$ since X_k is a covariate and X is the set of all covariates. However, the proposition tells us that we cannot assert that the absence of HTEs implies that X_k is not an MDV. The logic here resembles the previous case. Non-affine transformation $h(\cdot)$ violates Assumptions 1 and 2, which means that we cannot attribute a lack of heterogeneity in CATEs to a lack of differences in a given conditional AIEs. We purposely state a vacuous result to emphasize how little can be ascertained about mechanisms from the lack of HTEs when outcomes are transformed non-affinely.

Proposition 5. Suppose Assumptions 1 and 2 hold with respect to X_k for outcome Y. Let observed outcome $\tilde{Y}$ be a transformed outcome relative to Y. If HTEs do not exist with respect to X_k for outcome $\tilde{Y}$, then $X_k \in \mathrm{X}$.

Often, we make assumptions about the mapping h. For example, the mapping in Equation 13 imposes assumptions about how latent attitudes translate into Likert-scale responses. When we are willing to make such assumptions, we can refine Proposition 5 slightly. Specifically, in Proposition A6, we show that if Assumptions 1 and 2 hold for an absolutely continuous Y, which is transformed by Equation 13 (for any $Q \geq 2$ categories), if HTEs do not exist with respect to X_k, then one typical reason is that $X_k \notin \mathrm{X}^R$. Because $\mathrm{X}^{MDV} \subseteq \mathrm{X}^R$ we know then that $X_k \notin \mathrm{X}^{MDV}$ if $X_k \notin \mathrm{X}^R$. But as in Proposition 3 and Corollary 1, there are multiple possible explanations: our theory about how X_k relates to mechanism M_1 could be wrong or no

[15] Formally, if $X_k \in \mathrm{X}^R$, then there exist $x \neq x' \in X_k$ such that $Y(Z, M(Z; X_k = x, X_{\neg k}); X_k = x, X_{\neg k}) \neq Y(Z, M(Z; X_k = x', X_{\neg k}); X_k = x', X_{\neg k})$.

https://doi.org/10.1017/S0003055426101580 Published online by Cambridge University Press

TABLE 2. Summary of the Interpretation of Results

Case	Outcome type	HTEs?	What can be inferred about mechanism M_1?
1	Y	Yes	Difference in conditional indirect effect of mechanism M_1 is not zero $\Rightarrow$ Mechanism M_1 is *active* for at least one unit.
2	Y	No	Three possible explanations (not mutually exclusive): (a) Covariate X_k does not moderate the effect of mechanism M_1, regardless of whether the mechanism is active or not. (b) Mechanism M_1 is active and produces the same effect for all units. (No covariate moderates its indirect effect.) (c) Mechanism M_1 is inert (produces no effect) for all units. $\Rightarrow$*No information about activation* of mechanism *M* without further assumptions.
3	$\tilde{Y}$	Yes	Covariate X_k predicts outcome $\tilde{Y}$ $\Rightarrow$*No information about the activation* of mechanism M_1.
4	$\tilde{Y}$	No	Without further assumptions, provides no information about relationship between X_k and $\tilde{Y}$. *No information about the activation* of mechanism M_1.

Note: All cases assume that Assumptions 1 and 2 hold for covariate X_k, mechanism *M*, and outcome *Y*. Outcome $\tilde{Y}$ is a transformed outcome relative to *Y*.

MDV exists for mechanism M_1. These possibilities mean that we cannot make an inference about mechanism (non)-activation from the absence of HTEs with respect to X_k.

Summary of Results

In sum, our propositions characterize four cases into which we can classify attempts to ascertain mechanism activation from HTEs, as described in Table 2. This table suggests that in addition to the exclusion assumptions needed for identification of the difference of conditional average indirect effects of a mechanism (Assumptions 1 and 2), HTEs serve as a useful test of the activation of a mechanism when evaluated for an outcome that satisfies these exclusion assumptions. However, even with these outcomes, we can make an inference mechanism activation in the presence of heterogeneity, but not in the absence of heterogeneity. In the next section, we consider how these results should inform applied theories and research design.

APPLICATIONS AND GUIDANCE FOR RESEARCH DESIGN

How should these results guide future efforts to test mechanisms using HTEs? In this section, we use a set of four published applications to illustrate how these considerations should inform interpretation and prospective research design. In contrast to many methodological papers, our focus is on how theorized mechanisms underlying treatment effects link to a specific estimand: a difference in CATEs.[16] There are multiple ways to estimate this difference. For a randomized binary treatment, Z_i and binary moderator X_{ik}—a candidate MDV—the most straightforward estimator of this difference in CATEs is β_3 in the OLS regression equation:

$$Y_i = \beta_0 + \beta_1 Z_i + \beta_2 X_{ik} + \beta_3 Z_i X_{ik} + \varepsilon_i.$$

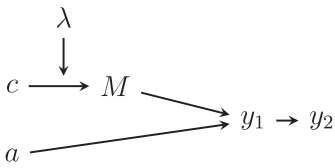
For guidance on estimation, we refer readers to excellent treatments of estimation of HTEs (or interaction effects), including Berry, DeMeritt, and Esarey (2009), Brambor, Clark, and Golder (2006), and Hainmueller, Mummolo, and Xu (2018). Additionally, see expositions of related inferential problems stemming from the limited statistical power of interaction effects (e.g., McClelland and Judd 1993) and the threat of multiple comparisons when HTEs in multiple moderators are estimated (e.g., Fink, McConnell, and Vollmer 2014; Gerber and Green 2012; Lee and Shaikh 2014).

In line with our motivating example, we select the four applications that examine empirically partisan alignment (or bias) as a possible moderator for the effect of corruption information about an incumbent candidate. Appendix A7.1 of the Supplementary Material describes the contexts and designs of these studies. Figure 3 provides a summary of the theories invoked in each study. Panel a—which is analogous to our running model (while omitting valence)—suggests that information about corruption gives way to an analogous distaste for corruption mechanism. But voters also value the partisanship of the incumbent. Distaste for corruption (the mechanism) and partisanship enter voter preferences for the incumbent as distinct and additively separable considerations of voters. Panel b is equivalent to Panel a, but includes a second mechanism linking corruption information to voter utility: a motivation of voters to coordinate votes within their precinct. Panel c returns to a single mechanism, voter distaste for corruption, but suggests that partisan alignment also conditions how voters process corruption information, suggesting the presence of motivated reasoning. Finally, Panel d includes only voter distaste for corruption and suggests that the degree to which voters are averse to corruption is a function of their partisan

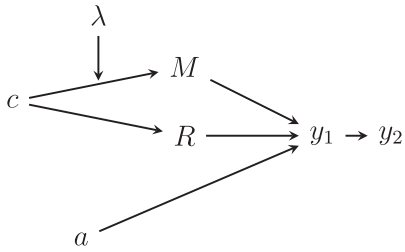
[16] These discussions generalize to a difference in conditional local average treatments (LATEs) or difference in CATEs on the treated (CATTs).

https://doi.org/10.1017/S0003055426101580 Published online by Cambridge University Press

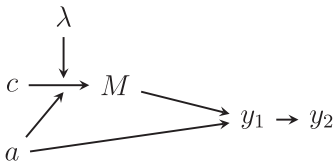
FIGURE 3. Four Theoretical Accounts of How Partisan Alignment (or Bias) and Information Relate to Voter Preferences for the Incumbent and Vote Choice



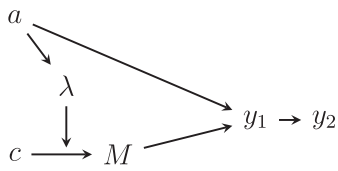
(a) Distaste for corruption (M) and partisan alignment (a) are distinct and additively separable arguments in voters' preferences for the incumbent (y_1), as in the running example.
Source: Eggers (2014)



(b) There are two mechanisms, distaste for corruption (M) and voter coordination motives (R). Partisan alignment (a) is distinct from both mechanisms and additively separable arguments in voters' preferences for the incumbent (y_1), as in the running example.
Source: Arias et al. (2019)



(c) Partisan alignment (bias) affects voter processing of corruption information via motivated reasoning and enters voters' preferences for the incumbent through a distinct channel.
Source: Anduiza, Gallego, and Muñoz (2013)



(d) Partisan alignment (a) conditions corruption aversion (λ), which moderates distaste for corruption (M). Alignment also enters voter's preferences for the incumbent (y_1) through a distinct channel.
Source: de Figueiredo, Hidalgo, and Kasahara (2023)

Note: *c* refers to corruption information; λ is a voter's corruption aversion; *M* is voters' distaste for observed corruption; *a* is voters' partisan bias toward the incumbent; y_1 is voter utility (preference) for the incumbent; and y_2 is vote choice for the incumbent.

leanings. We justify our representation of each cited theoretical account in Appendix A7.2 of the Supplementary Material. We note that the three papers with a single mechanism largely focus on voter distaste for observed corruption and two emphasize the role of partisanship in moderating this effect. In contrast, Arias et al. (2019) focus on the second mechanism—voter coordination—whereas we will focus (largely) on the distaste mechanism in our analysis. We note that, unlike our model, these papers do not focus on corruption aversion as a moderator, though such variation is generally consistent with the theories they articulate (see Appendix A7.2 of the Supplementary Material).

Our framework suggests four avenues for improved use of HTEs for mechanism detection which we explore below. First, it suggests a set of necessary theoretical considerations/arguments. Second, it suggests improvements in the interpretation of observed HTEs. Third, it offers recommendations for prospective research design. Finally, it provides guidance on when stronger theoretical assumptions may be useful to impose.

Three Essential Theoretical Considerations

Our framework identifies three attributes of a theory that are needed to support any analysis of causal mechanisms using HTE. First, researchers must *identify a set of candidate mechanisms.* The examples we cite are clear in positing a set of candidate mechanisms. Three—Panels a, c, and d—focus on a single mechanism: distaste for an incumbent's corruption. One—Panel b—instead suggests that distaste and voter coordination motives are two distinct mechanisms through which corruption

https://doi.org/10.1017/S0003055426101580 Published online by Cambridge University Press

information affects voter utility. The applications we discuss are quite clear about the mechanisms thought to underlie the effects that they measure.

Second, in order to use HTE to learn about a given mechanism, however, our analysis shows that one must (1) *invoke exclusion assumptions* and (2) *identify candidate MDVs.* In theories with a single mechanism—as in Panels a, c, and d of Figure 3—it is only necessary to defend Assumption 1 for a given covariate. In the case of the partisan alignment covariate, this means that we should justify the *absence* of a direct effect of corruption information that varies in partisan alignment. In Panel b, when there are multiple mechanisms, one would need to invoke Assumptions 1 and 2 for a given covariate, for example, partisan alignment. As represented in the papers (and therefore in Figure 3), the exclusion assumptions should hold for the voter utility outcome (y_1) in each case. We note that because these studies focus on a small (and well-articulated) set of mechanisms, assessing whether Assumptions 1 and (where relevant) 2 are plausible is relatively straightforward. In other types of studies that propose a larger set of candidate mechanisms, (1) additional exclusion assumptions must be invoked for each candidate mechanism and (2) justification of these exclusion assumptions becomes more difficult because there are more plausible violations of these assumptions.

The different theories vary in whether partisan alignment, a, should serve as an MDV for voter utility, y_1. In the first two theories, depicted in Panels a and b, partisan alignment is *not* a candidate MDV. We can see this because there is no arrow from a to the edge between c and M or the edge between M and y_1, or (in Panel b), the voter coordination mechanism (R). In contrast, a serves as an MDV for the distaste mechanism in Panels c and d. We can see this from the direct arrow from a to the edge between c and M in Panel c—which captures voters' motivated reasoning about the corruption disclosure—and from a to λ to the edge between c and M in Panel d, which captures the idea that partisan affiliation shapes corruption aversion, which in turn conditions the degree to which corruption information (c) generates a distaste for observed corruption.

Third, authors must consider the *relationship between outputs of the candidate mechanisms and measured outcomes.* The empirical studies associated with Panels a, b, and the field experiment in d measure aggregate vote share at the level of the constituency (Panel a) or precinct (Panels b and d). Within our framework, aggregate vote share can be thought of as $\sum_i y_{i2}/n$, where n is the number of voters.[17] This means that we should expect HTEs in partisan alignment for each of these studies. But recall that of these three studies, partisan alignment is only an MDV under the theory in d in which alignment conditions a voter's corruption aversion. In contrast, the survey-based studies in Panel c and the survey experiment component of Panel d measure outcomes that resemble voter utility through Likert scales. In both relevant Panels, the authors expect partisan affiliation to moderate the effect of corruption information on distaste for corruption (albeit for different reasons). So in contrast to our model (and the examples in Panels a and b), we would expect to observe HTE in partisan affiliation for these outcomes if the distaste for corruption mechanism is active.

Improving the Interpretation of HTEs as Mechanism Tests

By summarizing our results, Table 2 provides a guide for interpreting the presence or absence of HTEs as tests of a mechanism. Importantly, they point to an asymmetry in what we can infer about mechanisms from the presence versus absence of HTEs. Specifically, when exclusion assumptions hold, the presence of HTEs provides evidence that a mechanism is active. In contrast, the absence of HTEs does not rule out a mechanism or show that it is inert without further assumptions.

Our results are theoretical or identification results, meaning that they can be interpreted as what would obtain if we had an infinite sample. But, in practice, empiricists operate in a world with finite—and often relatively small—samples. This introduces statistical problems as well. In the context of HTEs, known limits to the statistical power for interaction effects (or terms) reduce our ability to 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 provides less informative about mechanism activation, 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.[18]

In the context of the studies that we examine, authors are admirably transparent about these limitations. For example, de Figueiredo, Hidalgo, and Kasahara (2023, 735) (Panel d) write "given the small samples, however, the difference between the two [CATE] estimates (the interaction) is not statistically significant. Still, the difference in magnitudes certainly suggests that [candidate's] voters are more sensitive to corruption-related information than supporters of other candidates." In sum, the combination of the asymmetry in the inferences about mechanisms with versus without HTEs combined with the limited ability to detect heterogeneity should give authors caution in the interpretation of HTEs as tests of mechanisms.

[17] In Panel a, Eggers (2014) also provides a survey-based measure of y_2 from the British Election Survey by analyzing self-reported vote choice. Some studies also consider vote share as a fraction of registered voters. This would require a model that allows voters to abstain, but yields substantively similar results.

[18] Selective reporting of significant results 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 probability that researchers infer that a mechanism is active when it is not.

https://doi.org/10.1017/S0003055426101580 Published online by Cambridge University Press

Guidance for Prospective 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.

Measure More Candidate MDVs

In terms of covariates, we are primarily concerned with which covariates are measured and the number of candidate MDVs (per mechanism) among those covariates. Following the exclusion assumptions, covariates are only useful for ascertaining mechanisms when (1) they are plausibly MDVs for a single mechanism and (2) they do not moderate direct effects. This observation suggests that special care must be taken when positing candidate MDVs. When pretreatment covariates are (largely) collected in baseline data collection, there is a need to posit MDVs and defend exclusion assumptions *ex-ante.* Such considerations require more theory and justification than are typically conveyed in the specification of moderation or HTE analyses in pre-analysis plans.

Further, it is very useful to have multiple candidate MDVs for a given mechanism. To see why, consider the case in which we have two candidate MDVs, X_1 and X_2 for mechanism M_1 and the exclusion assumptions hold for both candidate MDVs. Suppose that there do not exist HTEs in X_1 but there do exist HTEs in X_2. If we only measured HTEs with respect to X_1, following Proposition 3, we would not be able to ascertain whether the problem is with the theory ($X_1 \notin \mathbf{X}^{MDV}$) or whether there simply exist no MDV for mechanism M_1. If there exist HTEs in X_2, we can eliminate the possibility that there do not exist MDV for mechanism M_1. This would suggest that the theory with respect to X_1 is misspecified. This is useful insofar as it allows us to make an inference that mechanism M_1 is active. Note, however, that in order to leverage multiple candidate MDVs, the exclusion assumption must hold for each candidate MDV, which can be quite demanding.

The simplified presentation of the theories in Figure 3 is unlikely to reflect the full set of candidate MDVs for a given mechanism. However, we can explore this logic with respect to the theory in Panel c where there are two candidate MDVs. Specifically, in this study, corruption aversion and partisanship are posited to be MDVs for the distaste for corruption mechanism. Anduiza, Gallego, and Muñoz (2013) measure only the partisanship indicator and find that respondent distaste for a (hypothetical) incumbent's corruption varies in partisanship. Suppose instead that the authors had also measured corruption aversion—and detected HTEs in that variable—*while* failing to detect HTEs in partisanship. Such a finding would suggest that the distaste mechanism is active (via the HTEs in corruption aversion). It would additionally indicate that partisanship may not be an MDV for the distaste mechanism, perhaps casting doubt on the role of motivated reasoning in shaping a voters' distaste.

Prioritize Specific Outcomes for Mechanism Tests

Our focus on non-affine transformations of measured outcomes yields further recommendations for research design. If a goal of a research design is to distinguish between mechanisms or detect a posited mechanism, outcomes for which Assumptions 1 and 2 are plausible should be prioritized in HTE analysis. For example, if researchers had survey measures of voter utility from the incumbent and vote choice (from survey responses of administrative vote returns at the precinct level), any inference about the mechanism should be made on the basis of the utility measure. The logic behind this choice helps to convey the structure of the theorized mechanism and its relationship to the outcome measures. Moreover, ex-ante specification of the outcomes for which these assumptions are likely to hold can guide choices about which outcome variables to invest in measuring. This can also guide pre-analysis plans.

To be clear, our recommendation is *not* to avoid the estimation of HTEs for non-affinely transformed outcomes entirely. Indeed, in the study of elections, we typically care about vote choice more than voter utility, since this vote choice determines who wins office. But we should be clear about why HTEs on vote choice matter when they are not informative about mechanisms. They could be used to understand how to better target future corruption information interventions across the electorate (e.g., Athey and Wager 2021; Kitagawa and Tetenov 2018), to extrapolate effects to other electorates under a model (e.g., Egami and Hartman 2023), or to describe about the distributional impacts of the treatment. These are all valid—and perhaps underutilized—uses of HTEs which do not depend on the relationship between mechanisms and measured outcomes. We simply advise researchers to more carefully articulate the goals of their analysis.

Imposing Assumptions about Use of HTEs as Mechanism Tests

Our results suggest that the presence of HTEs is not informative of mechanism activation when the exclusion assumptions do not hold. Model-based approaches may be useful when researchers only have access to transformed outcomes or have reasons to doubt the validity of the exclusion assumptions. For example, for the field experimental and observational studies in Figure 3 (Panels a, b, and the field experiment in d), authors have vote share outcomes or self-reported vote choice, but recall that this is a transformed outcome. Furthermore, in Panel b, it may also be reasonable to believe that corruption aversion conditions both the distaste for corruption and voter coordination mechanisms. Can we make progress in these settings by imposing stronger assumptions about the mapping from voter preferences (utility) to vote choice?

We consider the possibility of invoking different models or sets of assumptions may aid in using HTEs to learn about mechanisms. Table 3 summarizes three problems identified by our analysis. Each problem is

https://doi.org/10.1017/S0003055426101580 Published online by Cambridge University Press

TABLE 3. Summary of Model- or Assumption-Based Alternatives to Exclusion Assumptions

	Problem	Model/assumptions	Conclusions
1	Exclusion assumptions fail because utility is transformed into a discrete choice.	Random utility model specifies a systematic and random component of utility that generates observed outcomes.	In settings in which mechanisms operate on utility but we observe choice, random-utility models can recover a model-based estimate of expected utility that can be used for analysis of mechanism activation through HTEs.
2	Exclusion assumptions fail because the outcome is transformed.	Assumption about monotonicity of CATEs in a candidate MDV X_k.	Monotonicity is not strong enough to provide information about mechanism activation without ancillary functional form assumptions.
3	Exclusion assumptions fail because the covariate may moderate multiple mechanisms.	Bayesian model of CATE magnitude under different mechanism activation profiles.	Plausible when we are willing to register priors about relative CATE magnitudes under different mechanisms.

paired with a statistical model or set of assumptions that we examine as a solution. The right column summarizes the conclusions of our analyses, which are detailed in greater depth in Appendix H of the Supplementary Material. In sum, two of the three models/sets of added assumptions are strong enough (in isolation) to allow HTE to provide information about mechanism activation, where they do not in the absence of these additional assumptions.

Modeling Non-Transformed Outcomes/Random Utility Models

In the context of transformations from utility to choice —like the transformation of voter utility from the incumbent into a vote for the incumbent—there exist widely-used random utility models that seek to recover preferences (utility) from choices. These models provide a functional mapping between an individual's utility (the outcome upon which the mechanism operates) and their choice by decomposing vote choice into an observed systematic and an unobserved random component. In our motivating model, the information treatment, the corruption aversion covariate, and a partisan alignment covariate are systematic and could, in principle, be observed. In contrast, the valence shock is random. In Panel b of Figure 3, corruption information (the treatment), corruption aversion, and the network structure of a precinct (or a sufficient statistic thereof) could be observed, but partisan alignment is random. By specifying the systematic component as a function of individual- or choice-specific covariates and assuming a distribution of the random component(s), researchers may be able to estimate the systematic component of utility.

Appendix A8.2 of the Supplementary Material analyzes the use of random-utility models in the context of HTE analysis. These models have two principal merits in the present context. First, the parameterization of the systematic component allows researchers to assess whether the exclusion assumptions are plausible under a given theoretical model. If these assumptions are plausible, the second benefit of a random utility model emerges. Specifically, it permits researchers to estimate (or "back out") an outcome for which HTEs can provide information about mechanism activation, specifically $\mathbb{E}[Y]$. Importantly, the invocation of a random-utility model is not free: it makes strong parametric assumptions about utility and its relationship to choice outcomes. Researchers may not be well-positioned to invoke or assess these assumptions. However, these assumptions allow for more formal examination of exclusion assumptions and may yield information which may permit learning about mechanism activation from HTEs.

While random utility models are the best established method for estimating actors' utility from choice, the broader approach could be useful for other types of transformed outcomes. Suppose that an empirical researcher believed that she observed $\tilde{Y} = h(Y)$ and that the exclusion assumptions held for Y (but not $\tilde{Y}$). If she were willing to impose some (invertible) functional form on $h(\cdot)$, it may be possible to evaluate (or estimate) $h^{-1}(\tilde{Y})$. As in the case of a random utility model, she could then conduct analysis with $E[Y]$.

Imposing Monotonicity

To this point, we analyze the current practice of examining the presence of HTEs as a test of mechanisms. However, examining the magnitudes of estimated CATEs may offer more information. We first consider whether the invocation of *monotonicity of CATEs*—a common empirical assumption (Manski 1997)—can provide evidence about mechanisms. In our context, monotonicity holds that for all $x' > x \in X_k$, $CATE(x') \geq (\leq) CATE(x)$. In the case of our model (and the applications in Panels a and b), corruption information should have a stronger (negative) effect on voter utility among voters with stronger corruption aversion (larger λ). Indeed, Remark 1 shows that for $\lambda > \lambda'$, $|CATE(y_1|X_1 = \lambda)| > |CATE(y_1|X_1 = \lambda')|$ for the voter utility outcome. Is it possible that this monotonicity in λ is maintained

https://doi.org/10.1017/S0003055426101580 Published online by Cambridge University Press

for the vote choice outcome? If this were the case, we could simply compare the magnitude of effects at different levels of the MDV (λ) for some suggestive evidence about mechanism activation. Similarly, could a violation of monotonicity of the form $\text{sign}(CATE(y_2|\lambda)) \neq \text{sign}(CATE(y_2|\lambda'))$ provide evidence against mechanism activation?

Unfortunately, in Appendix A8.1 of the Supplementary Material, we show that assuming monotonicity is not sufficient to provide information about mechanism activation through analysis of CATE magnitudes. Specifically, in Proposition A7 in the Supplementary Material, we show that monotonicity alone is not sufficient to ensure that HTEs take different signs when X_k is not an MDV. Moreover, it does not ensure that CATEs are monotonic in X_k even when X_k is an MDV and monotonicity is satisfied for the non-transformed outcome. These results show that we would need additional parametric assumptions for monotonicity to provide sufficient information to distinguish mechanism activation.

Learning from the Magnitude of HTEs

The magnitude of CATEs may provide additional information than the presence of HTE. For example, suppose that—in contrast to Panel b of Figure 3—theory suggested that the corruption aversion moderator was a candidate MDV for both the distaste and coordination mechanisms. If this were the case, the exclusion assumptions (Assumptions 1 and 2) would not be satisfied. However, we may have reasons to believe that the HTEs coming from one mechanism are large (in magnitude) while those coming from the other mechanism are small (in magnitude). In Appendix A8.3 of the Supplementary Material, we propose a simple Bayesian model that allows for inference about the activation of a specified mechanism from the magnitude of estimated effects. This model requires specification of priors over: (1) the probability of activation of a given mechanism and (2) the magnitude of HTEs under each candidate mechanism. Specification of these priors constitutes the invocation of additional assumptions.

We note that most theories in the social sciences admit directional—rather than point—predictions. By moving from the existence of HTE to their magnitude in this Bayesian setting, the model that we propose requires researchers to specify priors about the *size* of HTEs, not simply their existence or direction. While these priors allow us to glean some information about mechanisms when the exclusion assumption(s) are violated, more work is needed to guide researchers in specifying such priors from theories that are largely directional.

CONCLUSION

Social scientists routinely estimate HTEs with the goal of understanding which mechanisms generate treatment effects. By providing the first theoretical analysis of the relationship between HTEs and mechanisms, we show that detecting mechanisms with HTEs is far less straightforward than implied by current practice. Specifically, any link between a covariate (moderator) and a mechanism requires exclusion assumptions, so that the covariate does not moderate the effects of other mechanisms (or the direct effect). Even when these assumptions hold, we can only use HTEs to affirm the activation of a mechanism when (1) HTEs exist and (2) for transformed outcomes that preserve the additive separability of a mechanism's effect. 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 activation of a mechanism without stronger assumptions than those that we impose.

At present, mechanism detection is the modal use of HTEs in political science (see Table 1) and the modal method for mechanism detection (Blackwell, Ma, and Opacic 2024). However, mechanism detection is not the only use of HTE. Our results speak to contexts where mechanistic analysis is a goal. HTEs are also increasingly used for extrapolation of treatment effects to different populations/settings (Egami and Hartman 2023; Devaux and Egami 2022) and the targeting of treatments (Athey, Tibshirani, and Wager 2019; Kitagawa and Tetenov 2018). Our work does not directly speak to these uses of HTEs, because these methods do not seek to attribute observed effects to mechanisms (Slough and Tyson 2024).

Our analysis raises a number of issues and opportunities for future research to build upon. In particular, we emphasize that choices about outcome measures can complicate efforts to understand the substantive mechanisms at work. For example, even if the exclusion assumptions hold for one outcome, by imposing a common non-affine transformation on that outcome, the exclusion assumptions can be violated. This distinction has underappreciated implications for multiple quantitative methods to detect mechanism activation, including mediation analysis, analysis of treatment effects on intermediate outcomes, and efforts to link the sign of treatment effect to a (set of) mechanism(s). One potentially fruitful avenue for continued use of HTEs for mechanism detection would be to move from the presence of HTEs to their magnitude, as we outline in Appendix A8.3 of the Supplementary Material. Adoption of this approach will rely on the development of closer links between theoretical mechanisms and the size of reduced-form treatment effects than is current practice.

SUPPLEMENTARY MATERIAL

To view supplementary material for this article, please visit https://doi.org/10.1017/S0003055426101580.

DATA AVAILABILITY STATEMENT

Research documentation and data that support the findings of this study are openly available at the

https://doi.org/10.1017/S0003055426101580 Published online by Cambridge University Press

American Political Science Review Dataverse: https://doi.org/10.7910/DVN/G5EMO2.

ACKNOWLEDGEMENTS

We thank Scott Abramson, Neal Beck, Matt Blackwell, Matias Cattaneo, Andrew Eggers, Dorothy Kronick, Andrew Little, Justin Melnick, Molly Offer-Westort, Cyrus Samii, Rocío Titiunik, Scott Tyson, and Anna Wilke; two anonymous reviewers; seminar audiences at New York University, Princeton, and Berkeley; and participants at the NYU Abu Dhabi Theory in Methods Workshop and PolMeth XL for helpful feedback. We thank Carlo Cusumano for detailed feedback on the model and framework. T.S. is grateful for the support of a W. Glenn Campbell and Rita Ricardo-Campbell National Fellowship at the Hoover Institution.

CONFLICT OF INTEREST

The authors declare no ethical issues or conflicts of interest in this research.

ETHICAL STANDARDS

The authors affirm this research did not involve human participants.

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): 1008–20.

Acharya, Avidit, Matthew Blackwell, and Maya Sen. 2016. "Explaining Causal Findings without Bias: Detecting and Assessing Direct Effects." *American Political Science Review* 110 (3): 512–29.

Anduiza, Eva, Aina Gallego, and Jordi Muñoz. 2013. "Turning a Blind Eye: Experimental Evidence of Partisan Bias in Attitudes toward Corruption." *Comparative Political Studies* 46 (12): 1664–92.

Arias, Eric, Pablo Balán, Horacio Larreguy, John Marshall, and Pablo Querubín. 2019. "Information Provision, Voter Coordination, and Electoral Accountability: Evidence from Mexican Social Networks." *American Political Science Review* 113 (2): 475–98.

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

Ashworth, Scott, Christopher R. Berry, and Ethan Bueno de Mesquita. 2024. "Modeling Theories of Women's Underrepresentation in Elections." *American Journal of Political Science* 68 (1): 289–303.

Athey, Susan, Julie Tibshirani, and Stefan Wager. 2019. "Generalized Random Forests." *Annals of Statistics* 47 (2): 1148–78.

Athey, Susan, and Stefan Wager. 2021. "Policy Learning with Observational Data." *Econometrica* 89 (1): 133–61.

Berry, William D., Jacqueline H. R. DeMeritt, and Justin Esarey. 2009. "Testing for Interaction in Binary Logit and Probit Models: Is a Product Term Essential?" *American Journal of Political Science* 54 (1): 248–66.

Blackwell, Matthew, Ruofan Ma, and Aleksei Opacic. 2024. "Assumption Smuggling in Intermediate Outcome Tests of Causal Mechanisms." Preprint, arXiv:2407.07072v2.

Brambor, Thomas, William Roberts Clark, and Matt Golder. 2006. "Understanding Interaction Models: Improving Empirical Analyses." *Political Analysis* 14 (1): 63–82.

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): 375–91.

Bullock, John G., and Donald P. Green. 2021. "The Failings of Conventional Mediation Analysis and a Design-Based Alternative." *Advances in Methods and Practices in Psychological Science* 4 (4): 1–18.

de Figueiredo, Manuel F. P., F. Daniel Hidalgo, and Yuri Kasahara. 2023. "When Do Voters Punish Corrupt Politicians? Experimental Evidence from a Field and Survey Experiment." *British Journal of Political Science* 53 (2): 728–39.

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

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

Egami, Naoki, and Erin Hartman. 2023. "Elements of External Validity: Framework, Design, and Analysis." *American Political Science Review* 117 (3): 1070–88.

Eggers, Andrew C. 2014. "Partisanship and Electoral Accountability: Evidence from the UK Expenses Scandal." *Quarterly Journal of Political Science* 9 (4): 441–72.

Ferraz, Claudio, and Frederico Finan. 2008. "Exposing Corrupt Politicians: The Effects of Brazil's Publicly Released Audits on Electoral Outcomes." *Quarterly Journal of Economics* 123 (2): 703–45.

Fink, Günther, Margaret McConnell, and Sebastian Vollmer. 2014. "Testing for Heterogeneous Treatment Effects in Experimental Data: False Discovery Risks and Correction Procedures." *Journal of Development Effectiveness* 6 (1): 44–57.

Fu, Jiawei. 2024. "Extracting Mechanisms from Heterogeneous Effects: An Identification Strategy for Mediation Analysis." Preprint, arXiv:2403.04131.

Fu, Jiawei, and Tara Slough. 2026. "Replication Data for: Heterogeneous Treatment Effects and Causal Mechanisms." Harvard Dataverse. Dataset. https://doi.org/10.7910/DVN/G5EMO2.

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): 413–34.

Haim, Dotan, Nico Ravanilla, and Renard Sexton. 2021. "Sustained Government Engagement Improves Subsequent Pandemic Risk Reporting in Conflict Zones." *American Political Science Review* 115 (2): 717–24.

Hainmueller, Jens, Jonathan Mummolo, and Yiqing Xu. 2018. "How Much Should We Trust Estimates from Multiplicative Interaction Models? Simple Tools to Improve Empirical Practice." *Political Analysis* 27 (2): 163–92.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81 (396): 945–60.

Huang, Melody Y. 2024. "Sensitivity Analysis for the Generalization of Experimental Results." *Journal of the Royal Statistical Society Series A: Statistics in Society* 187 (4): 900–18.

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): 765–89.

Imai, Kosuke, Luke Keele, and Teppei Yamamoto. 2010. "Identification, Inference and Sensitivity Analysis for Causal Mediation Effects." *Statistical Science* 25 (1): 51–71.

https://doi.org/10.1017/S0003055426101580 Published online by Cambridge University Press

Imai, Kosuke, and Teppei Yamamoto. 2013. "Identification and Sensitivity Analysis for Multiple Causal Mechanisms: Revisiting Evidence from Framing Experiments." *Political Analysis* 21 (2): 141–71.
Incerti, Trevor. 2020. "Corruption Information and Vote Share: A Meta-Analysis and Lessons for Experimental Design." *American Political Science Review* 114 (3): 761–74.
Kitagawa, Toru, and Aleksey Tetenov. 2018. "Who Should Be Treated? Empirical Welfare Maximization Methods for Treatment Choice." *Econometrica* 86 (2): 591–616.
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 (4): 612–26.
Little, Andrew T., Keith E. Schnakenberg, and Ian R. Turner. 2022. "Motivated Reasoning and Democratic Accountability." *American Political Science Review* 116 (2): 751–67.
Manski, Charles F. 1997. "Monotone Treatment Response." *Econometrica* 65 (6): 1311–34.
McClelland, Gary H., and Charles M. Judd. 1993. "Statistical Difficulties of Detecting Interactions and Moderator Effects." *Psychological Bulletin* 114 (2): 376–90.
Moscowitz, Daniel. 2021. "Local News, Information, and the Nationalization of U.S. Elections." *American Political Science Review* 115 (1): 114–29.
Neyman, Jerzy. 1923. "Justification of Applications of the Calculus of Probabilities to the Solutions of Certain Questions in Agricultural Experimentation. Excerpts English Translation (1990)." *Statistical Science* 5 (4): 463–72.
Nilsson, Anton, Carl Bonander, Ulf Strömberg, and Jonas Björk. 2021. "A Directed Acyclic Graph for Interactions." *International Journal of Epidemiology* 50 (2): 613–9.
Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66 (5): 688.
Slough, Tara. 2023. "Phantom Counterfactuals." *American Journal of Political Science* 67 (1): 137–53.
Slough, Tara. 2024. "Bureaucratic Quality and Electoral Accountability." *American Political Science Review* 118 (4): 1931–50.
Slough, Tara, and Scott A. Tyson. 2023. "External Validity and Meta-Analysis." *American Journal of Political Science* 67 (2): 440–55.
Slough, Tara, and Scott A. Tyson. 2024. *External Validity and Evidence Accumulation*. New York: Cambridge University Press.
Slough, Tara, and Scott A. Tyson. 2025. "Sign-Congruence, External Validity, and Replication." *Political Analysis* 33 (3): 195–210.
Weinberg, Clarice R. 2007. "Can DAGs Clarify Effect Moderation?" *Epidemiology* 18 (5): 569–72.

https://doi.org/10.1017/S0003055426101580 Published online by Cambridge University Press