Heterogeneous Treatment Effects and Causal Mechanisms

Jiawei Fu and Tara Slough

March 2023

State of the Field

Credibility revolution $\longrightarrow$ use of research designs that facilitate identification and estimation of causal effects.

State of the Field

Credibility revolution $\longrightarrow$ use of research designs that facilitate identification and estimation of causal effects.

... but estimated causal effects do not (alone) tell us why or how

- Ultimately a question about causal mechanisms

State of the Field

Credibility revolution $\longrightarrow$ use of research designs that facilitate identification and estimation of causal effects.

... but estimated causal effects do not (alone) tell us why or how

- Ultimately a question about causal mechanisms

Various approaches (qualitative and quantitative) to evaluating mechanisms:

- Heterogeneous treatment effects (HTEs) estimated by treatment-by-covariate interactions is popular in applied work.

HTEs and Mechanisms: A Survey

We classify articles in three leading political science journals in 2021.

HTEs and Mechanisms: A Survey

We classify articles in three leading political science journals in 2021.

Journal (Issue)	Number of: Articles	Quant. articles	Pr(Report HTE| Quant. article)	Pr(Mechanism test| Report HTE)
AJPS (65)	61	41	0.56	0.87
APSR (115)	106	75	0.53	0.90
JoP (83)	142	106	0.55	0.83
Total	309	222	0.55	0.87

HTEs and Mechanisms: A Survey

We classify articles in three leading political science journals in 2021.

Journal (Issue)	Number of: Articles	Quant. articles	Pr(Report HTE| Quant. article)	Pr(Mechanism test| Report HTE)
AJPS (65)	61	41	0.56	0.87
APSR (115)	106	75	0.53	0.90
JoP (83)	142	106	0.55	0.83
Total	309	222	0.55	0.87

Takeaways:

1. Modal empirical article reports HTEs (treatment $\times$ covariate).
2. 87% of articles that report HTE use them to "test mechanisms."

Known vs. under-explored problems

Usual criticism of HTEs rests on statistical issues:

- Interactions are generally under-powered.
- Multiple comparisons problems (throwing spaghetti at the wall).

Known vs. under-explored problems

Usual criticism of HTEs rests on statistical issues:

- Interactions are generally under-powered.
- Multiple comparisons problems (throwing spaghetti at the wall).

We abstract from these problems by:

- Assuming an infinite sample.
- Looking at one covariate with specific relation to a mechanism.

Known vs. under-explored problems

Usual criticism of HTEs rests on statistical issues:

- Interactions are generally under-powered.
- Multiple comparisons problems (throwing spaghetti at the wall).

We abstract from these problems by:

- Assuming an infinite sample.
- Looking at one covariate with specific relation to a mechanism.

Under-explored:

Under what conditions do HTEs provide evidence of mechanism activation?

Outline

Motivating example: Exogenous shocks and voting behavior.

- Inspired by a model by Ashworth et al. (2018).
- Shows that HTE can emerge when posited mechanism is inert.

Outline

Motivating example: Exogenous shocks and voting behavior

Framework: We develop a framework to connect causal mechanisms to HTE with respect to covariates.

- Builds from causal mediation framework (Imai et al., 2010)
- New concepts, assumptions necessary for the HTE setting.

Outline

Motivating example: Exogenous shocks and voting behavior

Framework: We develop a framework to connect causal mechanisms to HTE with respect to covariates.

Results: What do we learn from the existence (or non-existence) of HTE with respect to covariates?

- For outcomes that are directly affected by the mechanism, HTE indicative of a mechanism under assumptions.
- For transformations of these directly affected outcomes, HTE are not necessarily indicative of a mechanism, even under these assumptions.

Outline

Motivating example: Exogenous shocks and voting behavior

Framework: We develop a framework to connect causal mechanisms to HTE with respect to covariates.

Results: What do we learn from the existence (or non-existence) of HTE with respect to covariates?

Discussion: Using these results to inform research design.

Motivating Example: Exogenous Shocks and Voting

Exogenous shocks and voting

Natural experiment on effect of an exogenous shock, ω, on voter behavior:

- A natural disaster (e.g., Healy and Malhotra, 2010; Huber et al., 2012)
- An economic crisis (e.g., Wolfers, 2002)
- A pandemic (e.g., Achen and Bartels, 2004; Baccini et al., 2021)

Exogenous shocks and voting

Natural experiment on effect of an exogenous shock, ω, on voter behavior:

- A natural disaster (e.g., Healy and Malhotra, 2010; Huber et al., 2012)
- An economic crisis (e.g., Wolfers, 2002)
- A pandemic (e.g., Achen and Bartels, 2004; Baccini et al., 2021)

Example: Ashworth et al., (2018):

1. Assume our adaption of model is true.
2. Suppose we could measure (some) model parameters directly.
    - Characterize causal estimands in terms of these parameters.
3. Ask: Can HTE provide evidence of voter learning mechanism?

Model

Incumbent at time of shock is of type $\theta \in \{\underline{\theta}, \overline{\theta}\}$, where $\overline{\theta} > \underline{\theta}$.

Model

Incumbent at time of shock is of type $\theta \in \{\underline{\theta}, \overline{\theta}\}$, where $\overline{\theta} > \underline{\theta}$.

Voters do not observe θ but may use governance outcome, g to update:

$$g = f(\theta, \omega) + \varepsilon.$$

- ω is increasing in the adversity of the shock
- ε is idiosyncratic shock drawn from symmetric, differentiable density, ϕ, that satisfies monotone likelihood ratio property relative to g.

Voter utility

Each voter's utility from a vote for politician, $p \in \{\mathcal{I}, \mathcal{C}\}$ is given by:

$$u_i^p = \theta^p + v_i \mathbb{1}(p = \mathcal{I})$$

Voter utility

Each voter's utility from a vote for politician, $p \in \{\mathcal{I}, \mathcal{C}\}$ is given by:

$$u_i^p = \theta^p + v_i \mathbb{I}(p = \mathcal{I})$$

Variation in the population of voters:

- $v_i \sim U(-1, 1)$ is a valence shock for the incumbent.
- Heterogeneous priors about incumbent: $\pi_i^{\mathcal{I}} \sim f_\pi$, $\pi_i^{\mathcal{I}} \in (0, 1)$.
- Common prior about the challenger: $\pi^{\mathcal{C}} \in (0, 1)$.

Sequence, voter behavior

Sequence:

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

Sequence, voter behavior

Sequence:

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

Voters' posteriors:

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

Sequence, voter behavior

Sequence:

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

Voters' posteriors:

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

A voter will vote for the incumbent if:

$$\underbrace{\beta(\overline{\theta}|\pi_i^I,\omega) + v_i}_{E[u_i^I]} \geq \underbrace{\pi^C}_{E[u_i^C]}$$

From theory to empirics

Treatment: Binary exposure to the shock $\omega \in \{\omega', \omega''\}$

From theory to empirics

Treatment: Binary exposure to the shock $\omega \in \{\omega', \omega''\}$

Outcomes: For the sake of exposition, consider two outcomes:

- Voter utility from the incumbent:

$$y_{1i} \equiv \beta(\overline{\theta}|\pi_i^I, \omega) + v_i$$

- Votes for the incumbent:

$$y_{2i} \equiv \mathbb{I}[\beta(\overline{\theta}|\pi_i^I, \omega) + v_i \geq \pi^C]$$

From theory to empirics

Treatment: Binary exposure to the shock $\omega \in \{\omega', \omega''\}$

Outcomes: For the sake of exposition, consider two outcomes:

- Voter utility from the incumbent:

$$y_{1i} \equiv \beta(\overline{\theta}|\pi_i^I, \omega) + v_i$$

- Votes for the incumbent:

$$y_{2i} \equiv \mathbb{1}[\beta(\overline{\theta}|\pi_i^I, \omega) + v_i \geq \pi^C]$$

Mechanism: Voter learning, not valence, since ω enters through voter's posterior.

Aside: DAG representation of interactions

“Interaction” effect representation not standardized

(Nilsson et al., 2020)

We need to be more precise in this talk.

Aside: DAG representation of interactions

"Interaction" effect representation not standardized

(Nilsson et al., 2020)

We need to be more precise in this talk.

The standard view

$$\begin{tikzcd} & X \arrow[d] \\ Z \arrow[r] & Y \end{tikzcd}$$

$$\frac{\partial Y}{\partial Z} \neq 0$$

$$\frac{\partial Y}{\partial X} \neq 0$$

$$\frac{\partial^2 Y}{\partial Z \partial X} \stackrel{?}{=} 0$$

Our notation

$$\begin{tikzcd} & X \arrow[d] \\ Z \arrow[r] & Y \end{tikzcd}$$

$$\frac{\partial Y}{\partial Z} \neq 0$$

$$\frac{\partial Y}{\partial X} \neq 0$$

$$\frac{\partial^2 Y}{\partial Z \partial X} = 0$$

$$\begin{tikzcd} & X \arrow[d] & \\ Z \arrow[rr] & {} & Y \end{tikzcd}$$

$$\frac{\partial Y}{\partial Z} \neq 0$$

$$\frac{\partial Y}{\partial X} \neq 0$$

$$\frac{\partial^2 Y}{\partial Z \partial X} \neq 0$$

The empiricist's question

Is the mechanism:

- Voter learning about I's type? $\longleftarrow$ the mechanism
- Amplification of I's valence? $\longleftarrow$ NOT the mechanism

The empiricist's question

Is the mechanism:

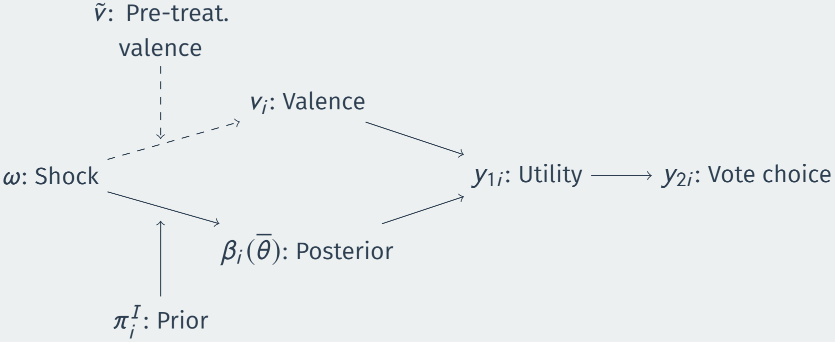
- Voter learning about I's type? $\longleftarrow$ the mechanism
- Amplification of I's valence? $\longleftarrow$ NOT the mechanism

$$
\begin{tikzcd}
 & & v_i\text{: Valence} \arrow[dr] & & \\
\omega\text{: Shock} \arrow[drr, ""{name=A}] & & & y_{1i}\text{: Utility} \arrow[r] & y_{2i}\text{: Vote choice} \\
 & & \beta_i(\overline{\theta})\text{: Posterior} \arrow[ur] & & \\
 & \pi_i^I\text{: Prior} \arrow[to=A] & & &
\end{tikzcd}
$$

The empiricist's question

Is the mechanism:

- Voter learning about I's type? ⟵ the mechanism
- Amplification of I's valence? ⟵ NOT the mechanism



Defining HTEs

To evaluate mechanisms, the empiricist will estimate CATEs for different levels of the (candidate) moderators: $x \in \{\pi_i^I, \tilde{v}\}$:

$$CATE(x') = E[y|\omega = \omega'', x = x'] - E[y|\omega = \omega', x = x']$$

Defining HTEs

To evaluate mechanisms, the empiricist will estimate CATEs for different levels of the (candidate) moderators: $x \in \{\pi_i^I, \tilde{v}\}$:

$$CATE(x') = E[y|\omega = \omega'', x = x'] - E[y|\omega = \omega', x = x']$$

There exist HTE in x if, for any $x' \neq x'' \in x$:

$$CATE(x'') - CATE(x') \neq 0.$$

Defining HTEs

To evaluate mechanisms, the empiricist will estimate CATEs for different levels of the (candidate) moderators: $x \in \{\pi_i^I, \tilde{v}\}$:

$$CATE(x') = E[y|\omega = \omega'', x = x'] - E[y|\omega = \omega', x = x']$$

There exist HTE in x if, for any $x' \neq x'' \in x$:

$$CATE(x'') - CATE(x') \neq 0.$$

We will evaluate the presence of HTE for:

- Outcomes: $y \in \{$Voter utility for I, Vote for $I\}$
- Potential moderators: $x \in \{$Prior belief about I, Pre-treatment valence$\}$

HTEs and mechanisms (results)

	y_1: Voter utility	y_2: Vote choice
x_1: Prior (π_i^I)	Mechanism HTE $CATE(\pi') \neq CATE(\pi'')$	
x_2: Valence ($\tilde{v}_i$)		

HTEs and mechanisms (results)

	y_1: Voter utility	y_2: Vote choice
x_1: Prior (π_i^I)	Mechanism HTE $CATE(\pi') \neq CATE(\pi'')$	
x_2: Valence ($\tilde{v}_i$)	Not a mechanism No HTE $CATE(\tilde{v}') = CATE(\tilde{v}'')$	

HTEs and mechanisms (results)

	y_1: Voter utility	y_2: Vote choice
x_1: Prior (π_i^I)	Mechanism HTE $CATE(\pi') \neq CATE(\pi'')$	Mechanism HTE $CATE(\pi') \neq CATE(\pi'')$
x_2: Valence ($\tilde{v}_i$)	Not a mechanism No HTE $CATE(\tilde{v}') = CATE(\tilde{v}'')$	

HTEs and mechanisms (results)

	y_1: Voter utility	y_2: Vote choice
x_1: Prior (π_i^I)	Mechanism HTE $CATE(\pi') \neq CATE(\pi'')$	Mechanism HTE $CATE(\pi') \neq CATE(\pi'')$
x_2: Valence ($\tilde{v}_i$)	Not a mechanism HTE $CATE(\tilde{v}') = CATE(\tilde{v}'')$	Not a mechanism HTE $CATE(\pi') \neq CATE(\pi'')$

HTEs and mechanisms: Implication/question

	y_1: Voter utility	y_2: Vote choice
x_1: Prior (π_i^I)	Mechanism HTE $CATE(\pi') \neq CATE(\pi'')$	Mechanism HTE $CATE(\pi') \neq CATE(\pi'')$
x_2: Valence ($\tilde{v}_i$)	Not a mechanism HTE $CATE(\tilde{v}') = CATE(\tilde{v}'')$	Not a mechanism HTE $CATE(\pi') \neq CATE(\pi'')$

HTE are not necessarily indicative of mechanism activation.

- To what extent is this general?

Framework

Set-up

A treatment, Z.

Set-up

A treatment, Z.

An outcome, Y.

- Continuous, directly observed variable of theoretical interest. Examples:
    - Utility (not choice)
    - Latent attitudes (not Likert responses)

Set-up

A treatment, Z.

An outcome, Y.

- Continuous, directly observed variable of theoretical interest. Examples:
  - Utility (not choice)
  - Latent attitudes (not Likert responses)

A set of pre-treatment covariates, $\mathbf{X}$.

Causal effects

Mediators as mechanism representations.

Causal effects

Mediators as mechanism representations.

Several causal effects typically described wrt causal mediation.

- Total effect (TE) of Z on Y.
- Indirect effect (IE_j) of Z on Y through mechanism j.
- Direct (unmediated) effect (DE) of Z on Y.

Causal effects

Mediators as mechanism representations.

Several causal effects typically described wrt causal mediation.

- Total effect (TE) of Z on Y.
- Indirect effect (IE_j) of Z on Y through mechanism j.
- Direct (unmediated) effect (DE) of Z on Y.

At the individual/unit level:

$$TE = DE + \sum_{j=1}^{J} IE_j$$

Causal effects

Mediators as mechanism representations.

Several causal effects typically described wrt causal mediation.

- Total effect (TE) of Z on Y.
- Indirect effect (IE_j) of Z on Y through mechanism j.
- Direct (unmediated) effect (DE) of Z on Y.

At the individual/unit level:

$$TE = DE + \sum_{j=1}^{J} IE_j$$

If a mechanism j is activated or present (for any unit), then there exists some unit for which $IE_j \neq 0$.

Estimands

Average treatment effect (ATE):

$$ATE = E[Y(z) - Y(z')] = E_X[DE + \sum_{j=1}^{J} IE_j]$$

Estimands

Average treatment effect (ATE):

$$ATE = E[Y(z) - Y(z')] = E_X[DE + \sum_{j=1}^{J} IE_j]$$

Conditional average treatment effects (CATE): Consider pre-treatment covariate $X_k \in X$. The CATE with respect to $X_k = x$ is:

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

Estimands

Average treatment effect (ATE):

$$ATE = E[Y(z) - Y(z')] = E_X[DE + \sum_{j=1}^{J} IE_j]$$

Conditional average treatment effects (CATE): Consider pre-treatment covariate $X_k \in X$. The CATE with respect to $X_k = x$ is:

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

Heterogeneous Treatment Effects (HTEs): HTEs exist with respect to pre-treatment covariate $X_k \in X$ iff:

$$CATE(X_k = x) \neq CATE(X_k = x')$$

for some $x \neq x' \in X_k$.

Reformulating the question

Original statement:

Under what conditions do HTEs provide evidence of mechanism activation?

Reformulating the question

Original statement:

Under what conditions do HTEs provide evidence of mechanism activation?

More precise version:

Under what conditions are HTEs with respect to X_k sufficient to show that there there exists some unit for which $IE_j \neq 0$?

Relationship to causal mediation

Mediation:

- Requires mediators to be measurable and measured.
- Assumes sequential ignorability.
- Seeks to estimate or bound IE_j and DE directly.

Relationship to causal mediation

Mediation:

- Requires mediators to be measurable and measured.
- Assumes sequential ignorability.
- Seeks to estimate or bound IE_j and DE directly.

Use of HTE:

- Does not require mediators to be measurable. (But we need specific measured covariates.)
- Invokes a set of exclusion assumptions.
- Seeks to demonstrate that $IE_j \neq 0$ for some unit.

HTEs and Mechanisms

Concept: Causal Indicator Variable (CIV)

Definition (Causal Indicator Variable)

Pre-treatment variable X_k is a causal indicator variable (CIV) for mechanism j if for some $x, x' \in X^k$, $IE_j(X_k = x) \neq IE_j(X_k = x')$.

Concept: Causal Indicator Variable (CIV)

Definition (Causal Indicator Variable)

Pre-treatment variable X_k is a causal indicator variable (CIV) for mechanism j if for some $x, x' \in X^k$, $IE_j(X_k = x) \neq IE_j(X_k = x')$.

$\mathbf{X}^{CIV}$ is the (possibly empty) set of covariates that satisfy definition.

Concept: Causal Indicator Variable (CIV)

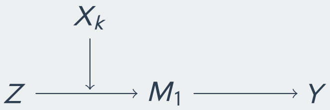
Definition (Causal Indicator Variable)

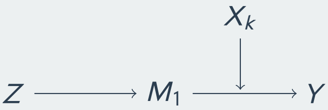
Pre-treatment variable X_k is a causal indicator variable (CIV) for mechanism j if for some $x, x' \in X^k$, $IE_j(X_k = x) \neq IE_j(X_k = x')$.

$\mathbf{X}^{CIV}$ is the (possibly empty) set of covariates that satisfy definition.

Two possibilities:

- $X_k \in X^{CIV}$ moderates the effect of treatment (Z) on mediator (M_j).
- $X_k \in X^{CIV}$ moderates the effect of the mediator (M_j) on outcome (Y).





Exclusion Assumption I

Assumption (Exclusion I)

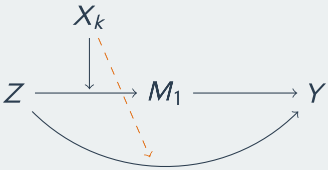
For any $x, x' \in X_k$, X_k *is non-linearly excluded to the direct effect such that* $DE(X_k = x) = DE(X_k = x')$.

Exclusion Assumption I

Assumption (Exclusion I)

For any $x, x' \in X_k$, X_k is non-linearly excluded to the direct effect such that $DE(X_k = x) = DE(X_k = x')$.

Direct effect of Z on Y cannot depend on X_k.



Exclusion Assumption II

Assumption (Exclusion II)

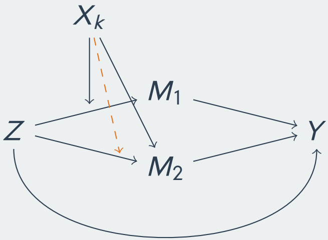
Given $z, z' \in Z$ *and* $x, x' \in X_k$, X_k *is non-linearly excluded to the indirect effect of mechanism* $j \neq 1$, IE_j, *if:* $IE_j(x) = IE_j(x')$.

Exclusion Assumption II

Assumption (Exclusion II)

Given $z, z' \in Z$ and $x, x' \in X_k$, X_k is non-linearly excluded to the indirect effect of mechanism $j \neq 1$, IE_j, if: $IE_j(x) = IE_j(x')$.

In other words, X_k is not a CIV for any other M_j.



HTEs as a test of the mechanism (#1 of 2)

Proposition

Suppose that Y is directly affected by mechanism 1 and Assumptions 1-2 hold with respect to X_k. If HTE exist with respect to X_k, then $X_k \in \boldsymbol{X}^{CIV}$ for mechanism 1.

HTEs as a test of the mechanism (#1 of 2)

Proposition

Suppose that Y is directly affected by mechanism 1 and Assumptions 1-2 hold with respect to X_k. If HTE exist with respect to X_k, then $X_k \in \boldsymbol{X}^{CIV}$ for mechanism 1.

Implication: By definition of CIV, HTE $\longrightarrow IE_1(X_k = x') \neq IE_1(X_k = x'')$ for some $x', x'' \in X_k$, which indicates that M_1 is active for some unit.

HTEs as a test of the mechanism (#1 of 2)

Proposition

Suppose that Y is directly affected by mechanism 1 and Assumptions 1-2 hold with respect to X_k. If HTE exist with respect to X_k, then $X_k \in \boldsymbol{X}^{CIV}$ for mechanism 1.

Implication: By definition of CIV, HTE $\longrightarrow IE_1(X_k = x') \neq IE_1(X_k = x'')$ for some $x', x'' \in X_k$, which indicates that M_1 is active for some unit.

Generalization: Holds for any non-zero linear transformation of Y, $L(Y)$.

HTEs as a test of the mechanism (#1 of 2)

Proposition

Suppose that Y is directly affected by mechanism 1 and Assumptions 1-2 hold with respect to X_k. If HTE exist with respect to X_k, then $X_k \in \boldsymbol{X}^{CIV}$ for mechanism 1.

Implication: By definition of CIV, HTE $\longrightarrow IE_1(X_k = x') \neq IE_1(X_k = x'')$ for some $x', x'' \in X_k$, which indicates that M_1 is active for some unit.

Generalization: Holds for any non-zero linear transformation of Y, $L(Y)$.

The usual logic for HTE, but note the assumptions.

- Implicit/unstated assumptions $\neq$ the absence of assumptions.

HTEs as test of mechanisms (#2 of 2)

Proposition

Suppose that Y is continuous and Assumptions 1 and 2 hold. If no HTE exist with respect to X_k, at least one of the following must true:

1. *Mechanism 1 is not active.*
2. $X_k \notin \boldsymbol{X}^{CIV}$

HTEs as test of mechanisms (#2 of 2)

Proposition

Suppose that Y is continuous and Assumptions 1 and 2 hold. If no HTE exist with respect to X_k, at least one of the following must true:

1. *Mechanism 1 is not active.*
2. $X_k \notin \boldsymbol{X}^{CIV}$

Absence of observed heterogeneity often equated with an inert mechanism.

HTEs as test of mechanisms (#2 of 2)

Proposition

Suppose that Y is continuous and Assumptions 1 and 2 hold. If no HTE exist with respect to X_k, at least one of the following must true:

1. *Mechanism 1 is not active.*
2. $X_k \notin \boldsymbol{X}^{CIV}$

Absence of observed heterogeneity often equated with an inert mechanism.

Alternate explanation: postulated relationship between X_k and M_1 is misspecified.

HTEs as test of mechanisms (#2 of 2)

Proposition

Suppose that Y is continuous and Assumptions 1 and 2 hold. If no HTE exist with respect to X_k, at least one of the following must true:

1. *Mechanism 1 is not active.*
2. $X_k \notin \boldsymbol{X}^{CIV}$

Absence of observed heterogeneity often equated with an inert mechanism.

Alternate explanation: postulated relationship between X_k and M_1 is misspecified.

Absence of HTE less informative than presence of HTE.

Summary (so far...)

Under Assumptions 1-2...

	Outcome variable is: Directly affected by M_1	Indirectly affected by M_1
$\exists$ HTE wrt X_k:	$X_k \in \mathbf{X}^{CIV}$ $\implies$ M_1 is active.	
$\nexists$ HTE wrt X_k:	$X_k \notin \mathbf{X}^{CIV}$ **or** M_1 not active	

Indirectly-affected outcomes

Why should we care:

Many attitudinal, behavioral outcomes are realizations of latent variables.

- Directly-affected outcomes are often unobserved by analyst.
- e.g., Vote choice vs. utility in our voting example.

Why should we care:

Many attitudinal, behavioral outcomes are realizations of latent variables.

- Directly-affected outcomes are often unobserved by analyst.
- e.g., Vote choice vs. utility in our voting example.

Distinction matters when:

- Observed indirectly-affected outcome is given by a non-linear transformation of the unobserved directly-affected outcome.
- Examples: models of (discrete) choice, Likert scales etc.

Why should we care:

Many attitudinal, behavioral outcomes are realizations of latent variables.

- Directly-affected outcomes are often unobserved by analyst.
- e.g., Vote choice vs. utility in our voting example.

Distinction matters when:

- Observed indirectly-affected outcome is given by a non-linear transformation of the unobserved directly-affected outcome.
- Examples: models of (discrete) choice, Likert scales etc.

Poses challenges for the quantitative evaluation of mechanisms.

One last concept

Useful to define $\mathbf{X}^R$ as the subset of measured covariates with a non-zero effect on directly-affected outcome Y. It is straightforward to see that:

$$\mathbf{X}^{CIV} \subseteq \mathbf{X}^R \subseteq \mathbf{X}$$

One last concept

Useful to define $\mathbf{X}^R$ as the subset of measured covariates with a non-zero effect on directly-affected outcome Y. It is straightforward to see that:

$$\mathbf{X}^{CIV} \subseteq \mathbf{X}^R \subseteq \mathbf{X}$$

In our motivating example, for the learning mechanism:

- $\mathbf{X}^{CIV} = \{\pi_i^I\}$
- $\mathbf{X}^R = \{\pi_i^I, \widetilde{v}_i\}$

HTEs on indirectly-affected outcomes (#1 of 2)

Proposition

Suppose that observed outcome $L(Y)$ is a non-linear transformation of directly-affected outcome Y and Assumptions 1 and 2 hold. If HTE exist with respect to X_k, then $X_k \in \boldsymbol{X}^R$.

HTEs on indirectly-affected outcomes (#1 of 2)

Proposition

Suppose that observed outcome $L(Y)$ is a non-linear transformation of directly-affected outcome Y and Assumptions 1 and 2 hold. If HTE exist with respect to X_k, then $X_k \in \mathbf{X}^R$.

Implication: Two possibilities:

- $X_k \in \mathbf{X}^{CIV} \implies M_1$ is active.
- $X_k \notin \mathbf{X}^{CIV} \implies M_1$ may or may not be active.

HTEs on indirectly-affected outcomes (#1 of 2)

Proposition

Suppose that observed outcome $L(Y)$ is a non-linear transformation of directly-affected outcome Y and Assumptions 1 and 2 hold. If HTE exist with respect to X_k, then $X_k \in \mathbf{X}^R$.

Implication: Two possibilities:

- $X_k \in \mathbf{X}^{CIV} \implies M_1$ is active.
- $X_k \notin \mathbf{X}^{CIV} \implies M_1$ may or may not be active.

Intuition: Using HTEs to detect mechanisms requires additive separability of X_k from DE and $IE_{j \neq 1}$.

- What Assumptions 1-2 buy us.
- Non-linear transformation $L(\cdot)$ does not preserve additive separability for indirectly-affected outcomes.

HTEs on indirectly-affected outcomes (#2 of 2)

Proposition

Suppose that observed outcome $L(Y)$ is a non-linear mapping of directly-affected outcome Y and Assumptions 1 and 2 hold. If HTE do not exist with respect to X_k, then $X_k \notin \boldsymbol{X}^R$.

HTEs on indirectly-affected outcomes (#2 of 2)

Proposition

Suppose that observed outcome $L(Y)$ is a non-linear mapping of directly-affected outcome Y and Assumptions 1 and 2 hold. If HTE do not exist with respect to X_k, then $X_k \notin \mathbf{X}^R$.

Implication: We know that $X_k \notin \mathbf{X}^{CIV} \implies M_1$ may or may not be active.

- So no information about mechanism activation.

Summary

Under Assumptions 1-2...

	Outcome variable is: Directly affected by M_1	Indirectly affected by M_1
$\exists$ HTE wrt X_k:	$X_k \in \mathbf{X}^{CIV}$ $\implies$ M_1 is active.	$X_k \in \mathbf{X}^{R}$ M_1 active or inactive
$\nexists$ HTE wrt X_k:	$X_k \notin \mathbf{X}^{CIV}$ **or** M_1 not active	$X_k \notin \mathbf{X}^{R}$ M_1 active or inactive

Is this really an issue?

We cannot know in any specific case whether heterogeneity comes from causal heterogeneity or the transformation to the observed outcome.

Is this really an issue?

We cannot know in any specific case whether heterogeneity comes from causal heterogeneity or the transformation to the observed outcome.

But, we can show via simulation in which we control the DGP that it is very easy to generate these dynamics.

Example: Support for greenhouse gas regulation

Tangentially inspired by setup of Little et al. (2021).

Example: Support for greenhouse gas regulation

Tangentially inspired by setup of Little et al. (2021).

Suppose we randomly assign partisan voters (in the US) to information about dangers of greenhouse gases (GHGs). They update beliefs via two mechanisms:

- Accuracy motivates
- Directional motives

Example: Support for greenhouse gas regulation

Tangentially inspired by setup of Little et al. (2021).

Suppose we randomly assign partisan voters (in the US) to information about dangers of greenhouse gases (GHGs). They update beliefs via two mechanisms:

- Accuracy motivates
- Directional motives

HTE idea: Partisanship moderates only directional motives.

Example: Support for greenhouse gas regulation

Tangentially inspired by setup of Little et al. (2021).

Suppose we randomly assign partisan voters (in the US) to information about dangers of greenhouse gases (GHGs). They update beliefs via two mechanisms:

- Accuracy motivates
- Directional motives

HTE idea: Partisanship moderates only directional motives.

Outcome: Binary indicator for "favors increased GHG regulation," assumed to be increasing in dangers of GHGs.

Simulation with real data

ANES 2020 data among declared Democrats and Republicans ($n = 2,883$).

- 82.2% of D's and 38.3% of R's support increased GHG regulation.

Simulation with real data

ANES 2020 data among declared Democrats and Republicans ($n = 2,883$).

- 82.2% of D's and 38.3% of R's support increased GHG regulation.

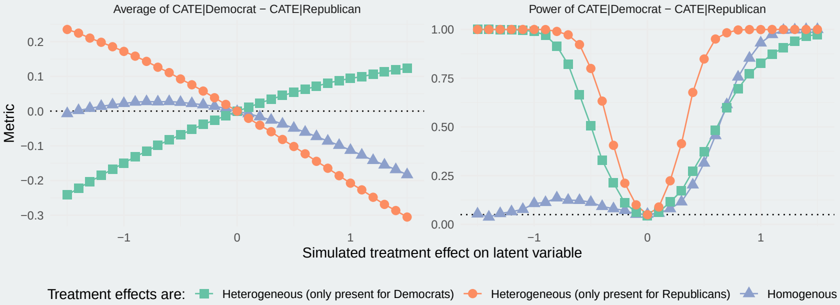
Simulation:

1. Using observed support and matrix of demographic covariates to predict underlying belief in dangers of GHGs. This will be $Y_i(0)$.
2. Simulate treated potential outcomes $Y_i(1)$ by adding treatment effect τ:

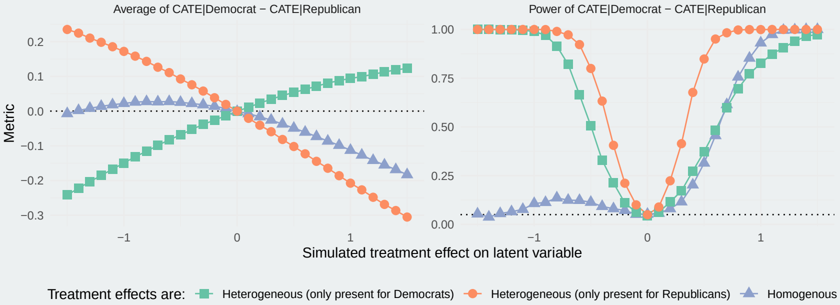
$$Y_i(1) = Y_i(0) + \tau \mathbb{1}(\text{Partisanship}_i = P)$$

3. Randomly assign treatment to half the sample to reveal $Y_i(Z)$.
4. Reveal observed outcome $\widetilde{Y}_i \sim \text{Bernoulli}(\text{logit}^{-1}(Y_i(Z)))$.
5. Estimate treatment effects on observed outcome $\widetilde{Y}_i$.

Simulation results



Simulation results



We observe HTE in partisanship for all $\tau \neq 0$ even when treatment effects (on latent variable) are homogenous!

- Magnitude and sign depend on density on the latent variable.

Discussion: Implications for Research Design

Improving the use of HTE: Role of theory

We need more explicit theory to use HTE for mechanism detection.

Improving the use of HTE: Role of theory

We need more explicit theory to use HTE for mechanism detection.

Three central questions:

1. What are the candidate mechanisms?
2. What is the relationship between a given covariate, X_k, and each of the candidate mechanisms?
    - For which mechanism (j), is X_k a candidate mechanisms?
    - Are exclusion assumptions plausible for other mechanisms?
3. Do mechanisms directly affect measured outcomes?
    - If so, which outcomes?

Improving the use of HTE: Better research design

Prioritizing different outcomes: Can we measure more directly-affected outcomes?

- When we have the ability to measure directly-affected outcomes, we should do so.
- Possibly more latent-variable estimation for outcomes—mixed feelings here.

Improving the use of HTE: Better research design

Prioritizing different outcomes: Can we measure more directly-affected outcomes?

- When we have the ability to measure directly-affected outcomes, we should do so.
- Possibly more latent-variable estimation for outcomes—mixed feelings here.

Which covariates should be measured?

- Posit $\mathbf{X}^{CIV}$ before data collection
- Possible implementation: map covariates to candidate mechanisms in pre-analysis plans

Improving the interpretation of HTE

Interpret lack of HTE accurately:

- Does not "rule out" a candidate mechanism (or show that it is inert), even when we have a directly-observed outcome.

Improving the interpretation of HTE

Interpret lack of HTE accurately:

- Does not "rule out" a candidate mechanism (or show that it is inert), even when we have a directly-observed outcome.

Consider implications of lack of power for inferences about the mechanism.

- Absent p-hacking/publication bias etc., low power $\longrightarrow \downarrow$ ability to detect HTE.
- But if lack of HTE has two sources (inert mechanism or misspecified theory), this provides less information.

Can we assume more?

Suppose we only observe an indirectly-affected outcome (i.e., vote choice).

Can we assume more?

Suppose we only observe an indirectly-affected outcome (i.e., vote choice).

Monotonicity seems like a useful assumption, e.g.:

$$CATE(X_k = x') > (<) CATE(X_k = x) \text{ for any } x' > x \in X_k.$$

Can we assume more?

Suppose we only observe an indirectly-affected outcome (i.e., vote choice).

Monotonicity seems like a useful assumption, e.g.:

$$CATE(X_k = x') > (<) CATE(X_k = x) \text{ for any } x' > x \in X_k.$$

Can we use an assumption of monotonicity to link HTEs to mechanisms?

- In general, not absent further assumptions on the distribution of the directly observed outcome, Y.
- Could impose monotonicity + distributional assumption on Y to facilitate inference about mechanisms.

Thank you!