Heterogeneous Treatment Effects and Causal Mechanisms

Jiawei Fu and Tara Slough—NYU

February 2023

estimation of causal effects.

State of the field

Credibility revolution \rightarrow 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 \rightarrow 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 x covariate interactions is very popular in applied work. HTE and mechanisms: a survey

We classify all articles published in three leading political science journals in 2021:

HTE and mechanisms: a survey

We classify all articles published in three leading political science journals in 2021:

	Number of:		Pr(Report HTE)	Pr(Mechanism test)
Journal (Issue)	Articles	Quant. articles	Quant. article)	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

HTE and mechanisms: a survey

We classify all articles published in three leading political science journals in 2021:

	Number of:		Pr(Report HTE)	Pr(Mechanism test)
Journal (Issue)	Articles	Quant. articles	Quant. article)	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 underpowered.
- Multiple comparisons problems (throwing spaghetti at the wall).

Known vs. under-explored problems

Usual criticism of HTEs rests on statistical issues:

- Interactions are generally underpowered.
- 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 mechanisms.

Known vs. under-explored problems

Usual criticism of HTEs rests on statistical issues:

- Interactions are generally underpowered.
- 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 mechanisms.

Under-explored problem:

Under what conditions do HTEs provide evidence of mechanism activation?

Motivating example: Exogenous shocks and voting behavior

- Based on model by Ashworth et al. (2018).
- Shows that HTE can emerge when relevant mechanism is inert.

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 concept, assumptions necessary for the HTE setting.

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 continuous covariates of theoretical interest, HTE indicative of a mechanism under assumptions.
- For realizations of latent theoretical constructs, HTE are not necessarily indicative of a mechanism, even under these assumptions.

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:

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

Example: Ashworth, Bueno de Mesquita, Friedenberg (2018):

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

Model set-up

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

Model set-up

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:

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

- $\circ \omega$ 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, $p \in \{I, C\}$ is given by:

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

Voter utility

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

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

Variation in the population of voters:

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

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:

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

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:

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

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:

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

Voter votes for the incumbent:

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

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:

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

Voter votes for the incumbent:

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

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

Aside: DAG representation of interactions

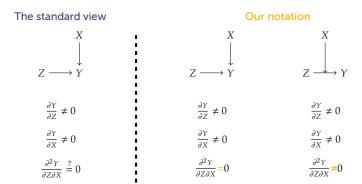
Representation of "interaction" effects in DAGs is not standard (Nilsson et al., 2020)

• We need to be a bit more precise in this talk

Aside: DAG representation of interactions

Representation of "interaction" effects in DAGs is not standard (Nilsson et al., 2020)

• We need to be a bit more precise in this talk



The empiricist's question

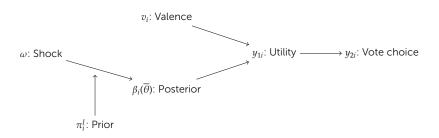
Is the mechanism:

- Voter learning about I's type? ← the mechanism
- $\circ \ \ \, \text{Amplication of I's valence?} \leftarrow \text{NOT the mechanism}$

The empiricist's question

Is the mechanism:

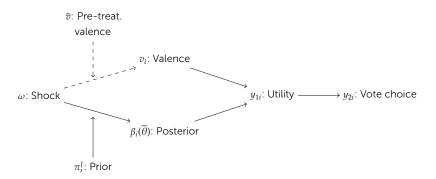
- Voter learning about I's type? ← the mechanism
- Amplication of I's valence? ← NOT the mechanism



The empiricist's question

Is the mechanism:

- \circ Voter learning about I's type? \leftarrow the mechanism
- $\circ \ \ \, \text{Amplication of I's valence?} \leftarrow \text{NOT the mechanism}$



Defining HTEs

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

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

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

HTEs and mechanisms (results)

	y_1 : Voter utility	y ₂ : Vote choice
	Mechanism	
x_1 : Prior (π_i^I)	HTE	
	$CATE(\pi') \neq CATE(\pi'')$	
x_2 : Valence (\tilde{v}_i)		

	y_1 : Voter utility	y ₂ : Vote choice
	Mechanism	
x_1 : Prior (π_i^I)	HTE	
	$CATE(\pi') \neq CATE(\pi'')$	
x_2 : Valence (\tilde{v}_i)	Not a mechanism	
	No HTE	
	$CATE(\tilde{v}') = CATE(\tilde{v}'')$	

	y_1 : Voter utility	y ₂ : Vote choice
	Mechanism	Mechanism
x_1 : Prior (π_i^I)	HTE	HTE
	$CATE(\pi') \neq CATE(\pi'')$	$CATE(\pi') \neq CATE(\pi'')$
	Not a mechanism	
x_2 : Valence (\tilde{v}_i)	No HTE	
	$CATE(\tilde{v}') = CATE(\tilde{v}'')$	

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

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

HTE are not necessarily indicative of mechanism activation.

• To what extent is this general?



Three main components

A treatment, Z.

An outcome, Y.

- Continuous, we directly observe the variable of theoretical interest.
- (Latent variable case later.)

A set of pre-treatment covariates, X.

Causal Effects

Mediators as mechanism representations.

Several causal effects typically described wrt causal mediation.

- Total effect (TE) of Z on Y.
- ∘ Indirect effect (IE_i) of Z on Y through mediator/mechanism $j \in \{1,2\}$.
 - Mechanism of interest will be j = 1.
 - Composite of all other mechanisms is j = 2.
- Direct (unmediated) effect (DE) of Z on Y.

At the individual/unit level:

$$TE = DE + IE_1 + IE_2$$

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

Estimands

Average treatment effect (ATE):

$$ATE = E_X[Y(z) - Y(z')]$$
$$= E_X[DE + IE_1 + IE_2]$$

Estimands

Average treatment effect (ATE):

$$ATE = E_X[Y(z) - Y(z')]$$
$$= E_X[DE + IE_1 + IE_2]$$

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):

$$ATE = E_X[Y(z) - Y(z')]$$
$$= E_X[DE + IE_1 + IE_2]$$

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 mediation

Mediation is advocated as a method for quantitative evaluation of mechanisms.

Relationship to mediation

Mediation is advocated as a method for quantitative evaluation of mechanisms.

Mediation.

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

Relationship to mediation

Mediation is advocated as a method for quantitative evaluation of mechanisms.

Mediation:

- Requires mediators to be measurable and measured.
- Assumes sequential ignorability.
- Seeks to estimate or bound *IE*_i 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 $IE_i \neq 0$ 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 1 if for some $x, x' \in X^k$, $IE_1(X_k = x) \neq IE_1(X_k = x')$.

Concept: Causal Indicator Variable (CIV)

Definition (Causal Indicator Variable)

Pre-treatment variable X_k is a causal indicator variable (CIV) for mechanism 1 if for some $x, x' \in X^k$, $IE_1(X_k = x) \neq IE_1(X_k = x')$.

 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 1 if for some $x, x' \in X^k$, $IE_1(X_k = x) \neq IE_1(X_k = x')$.

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

Two possibilities:

- ∘ $X_k ∈ X^{CIV}$ moderates the effect of treatment (Z) on mediator (M_i) .
- ∘ $X_k ∈ X^{CIV}$ moderates the effect of the mediator (M_i) on outcome (Y).

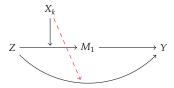


Exclusion assumption I

Assumption (Exclusion I)

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

• Direct effect of Z on Y cannot depend on X_k .

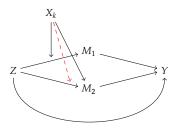


Exclusion assumption II

Assumption (Exclusion II)

For any $x, x' \in X_k$, X_k is irrelevant to the indirect effect of M_2 such that $IE_2(X_k = x) = IE_2(X_k = x')$.

• In other words, X_k is not a CIV for M_2 .



HTEs as test of mechanisms (#1 of 2)

Proposition

Suppose that Y is continuous and Assumptions 1 and 2 hold. If HTE exist with respect to X_k , then $X_k \in X^{CIV}$.

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

The usual logic for HTE, but note the assumptions.

• Implicit/unstated assumptions ≠ 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. $X_k \notin X^{CIV}$
- 2 Mechanism 1 is not active.

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 (with a conjecture)

Under Assumptions 1 and 2:

	Observed, continuous Y	Latent Y
3 HTE:	$X_k \in X^{CIV} \implies$	
	M_1 active	
∄ HTE:	$X_k \notin X^{CIV}$ or	
	M_1 not active	



Why should we care?

Many attitudinal, behavioral outcomes are realizations of latent variables.

- Example: decisions based on utility maximization.
- Vote choice vs. utility in our motivating example

Why should we care?

Many attitudinal, behavioral outcomes are realizations of latent variables.

- Example: decisions based on utility maximization.
- Vote choice vs. utility in our motivating example

Matters whenever mechanisms act upon the latent variable.

- Much of the time
- In the example, voter learning changes expected utility

Why should we care?

Many attitudinal, behavioral outcomes are realizations of latent variables.

- Example: decisions based on utility maximization.
- Vote choice vs. utility in our motivating example

Matters whenever mechanisms act upon the latent variable.

- Much of the time
- In the example, voter learning changes expected utility

Poses challenges for the detection of mechanisms

• Through HTEs and likely other approaches.

Additional structure, concept

Suppose that Y^* is the latent variable. We observe $Y = g(Y^*)$.

- Concern emerges when $g(\cdot)$ is non-linear.
- e.g., Discrete outcomes (choices, attitudinal scales etc.)

Additional structure, concept

Suppose that Y^* is the latent variable. We observe $Y = g(Y^*)$.

- Concern emerges when $g(\cdot)$ is non-linear.
- e.g., Discrete outcomes (choices, attitudinal scales etc.)

Useful to define X^{Rel} as the subset of measured covariates with a non-zero effect on latent outcome Y^* . It is straightforward to see that:

$$X^{CIV}\subseteq X^{Rel}\subseteq X$$

Additional structure, concept

Suppose that Y^* is the latent variable. We observe $Y = g(Y^*)$.

- Concern emerges when $g(\cdot)$ is non-linear.
- e.g., Discrete outcomes (choices, attitudinal scales etc.)

Useful to define X^{Rel} as the subset of measured covariates with a non-zero effect on latent outcome Y^* . It is straightforward to see that:

$$X^{CIV} \subseteq X^{Rel} \subseteq X$$

In our motivating example, for the learning mechanism:

- $\circ \quad X^{CIV} = \{\pi_i^I\}$
- $\circ \quad X^{Rel} = \{\pi_i^I, v_i\}$

Intuition and result

Mechanism test relies on additive separability of X_k from DE and IE_2 on the latent variable.

- What Assumptions 1-2 buy us.
- But a non-linear mapping from the latent implies observed outcome does not preserve additive separability on observed outcomes.

Proposition

Suppose that Y is the observed realization of the latent variable Y^* and Assumptions 1 and 2 hold. If HTE exist with respect to X_k , then $X_k \in X^{Rel}$.

Implication: Two possibilities:

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

Jiawei Fu and Tara Slough—NYU Latent Outcomes

Summary

Under Assumptions 1 and 2:

	Observed, continuous Y	Latent Y
∃ HTE:	$X_k \in X^{CIV} \implies$	$X_k \in X^{Rel}$
	M_1 active	M_1 active or not active
∄ HTE:	$X_k \notin X^{CIV}$ or	
	M_1 not active	

Summary

Under Assumptions 1 and 2:

	Observed, continuous Y	Latent Y
∃ HTE:	$X_k \in X^{CIV} \implies$	$X_k \in X^{Rel}$
	M_1 active	M_1 active or not active
∄ HTE:	$X_k \notin X^{CIV}$ or	$(X_k \notin X^{Rel})$
	M_1 not active	M_1 active or not active)

Discussion

More explicit theory in applied empirical work:

- What are the candidate mechanisms?
- How does covariate X_k relate theoretically to these mechanisms?
 - Is X_k a possible CIV for a mechanism?
 - Is Assumption 2 plausible with respect to other mechanisms?
- What from the theory is observed vs. latent?

More explicit theory in applied empirical work.

More accurate interpretation of findings:

- If willing to make Assumptions #1 and #2, what could have been learned?
- e.g., Lack of (detected) HTE does not imply mechanism inactive.

More explicit theory in applied empirical work.

More accurate interpretation of findings.

Better design of studies:

- \circ What variables are (theoretically) most likely to be in X^{CIV} ? \leftarrow collect them!
- Can latent outcomes be avoided? Can we measure theoretical constructs more directly?

More explicit theory in applied empirical work.

More accurate interpretation of findings.

Better design of studies.

Alternate uses of HTE (besides mechanism tests):

- e.g., Targeting of subsequent interventions.
- Assessment of degree of heterogeneity in treatment effects.

Take-aways

- 1. Using HTEs to test for activation of a mechanism:
 - Invokes assumptions about the covariate and other possible mechanisms.
 - May not be informative even under these assumptions.
- 2. Analysis of theoretical issues complements statistical critiques.
 - e.g., Concerns about power assume we could learn something if we had more power.

www.taraslough.com

Thank you!