Survey Experiments in Practice

Thomas J. Leeper

Government Department
London School of Economics and Political Science

18 January 2017

Activity!

Activity!

1. Ask you to guess a number

Activity!

1. Ask you to guess a number
2. Number off 1 and 2 across the room

Activity!

1. Ask you to guess a number
2. Number off 1 and 2 across the room
3. Group 2, close your eyes

Activity!

Group 1

Think about whether the population of Chicago is more or less than 500,000 people. What do you think the population of Chicago is?

Activity!

1. Ask you to guess a number
2. Number off 1 and 2 across the room
3. Group 2, close your eyes
4. Group 1, close your eyes

Activity!

Group 2

Think about whether the population of Chicago is more or less than 10,000,000 people. What do you think the population of Chicago is?

Enter your data

- Go here: `http://bit.ly/297vEdd`
- Enter your guess and your group number

Results

- True population: 2.79 million

Results

- True population: 2.79 million
- What did you guess? (See Responses)

Results

- True population: 2.79 million
- What did you guess? (See Responses)
- What's going on here?
  - An experiment!
  - Demonstrates "anchoring" heuristic

Results

- True population: 2.79 million
- What did you guess? (See Responses)
- What's going on here?
  - An experiment!
  - Demonstrates "anchoring" heuristic
- Experiments are easy to analyze, but only if designed and implemented well

1. History and Logic of Experiments
2. From Theory to Design
3. Operationalization Principles
   - Common Paradigms and Examples
4. Sources of Heterogeneity
   - Settings
   - Unit
   - Treatments
   - Outcomes
5. Beyond One-Shot Designs
6. Presentations/Conclusion

Who am I?

- Thomas Leeper
- Assistant Professor in Political Behaviour at London School of Economics
    - 2013–15: Aarhus University (Denmark)
    - 2008–12: PhD from Northwestern University (Chicago, USA)
    - Birth–2008: Minnesota, USA
- Interested in public opinion and political psychology
- Email: t.leeper@lse.ac.uk

Who are you?

- Where are you from?
- Have you designed a survey and/or experiment before?
- What do you hope to learn from the course?

Quick Survey

Quick Survey

1. How many of you have worked with survey data before?

Quick Survey

1. How many of you have worked with survey data before?
2. Of those, how many of you have *performed* a survey before?

Quick Survey

1. How many of you have worked with survey data before?
2. Of those, how many of you have *performed* a survey before?
3. How many of you have worked with experimental data before?

Quick Survey

1. How many of you have worked with survey data before?
2. Of those, how many of you have *performed* a survey before?
3. How many of you have worked with experimental data before?
4. Of those, how many of you have *performed* an experiment before?

Course Materials

All material for the course is available at:

`http://www.thomasleeper.com/surveyexpcourse/`

Learning Outcomes

By the end of the day, you should be able to...

Learning Outcomes

By the end of the day, you should be able to...

1. Explain how to analyze experiments quantitatively.

Learning Outcomes

By the end of the day, you should be able to...

1. Explain how to analyze experiments quantitatively.
2. Explain how to design experiments that speak to relevant research questions and theories.

Learning Outcomes

By the end of the day, you should be able to...

1. Explain how to analyze experiments quantitatively.
2. Explain how to design experiments that speak to relevant research questions and theories.
3. Evaluate the uses and limitations of several common survey experimental paradigms.

Learning Outcomes

By the end of the day, you should be able to...

1. Explain how to analyze experiments quantitatively.
2. Explain how to design experiments that speak to relevant research questions and theories.
3. Evaluate the uses and limitations of several common survey experimental paradigms.
4. Identify practical issues that arise in the implementation of experiments and evaluate how to anticipate and respond to them.

Experiments: Definition

Oxford English Dictionary defines "experiment" as:

1. A scientific procedure undertaken to make a discovery, test a hypothesis, or demonstrate a known fact
2. A course of action tentatively adopted without being sure of the outcome

Experiments: History

- "Experiments" have a very long history
- Major advances in design and analysis of experiments based on agricultural and later biostatistical research in the 19th century (Fisher, Neyman, Pearson, etc.)
- First randomized, controlled trial (RCT) by Peirce and Jastrow in 1884

Experiments: History

- "Experiments" have a very long history
- Major advances in design and analysis of experiments based on agricultural and later biostatistical research in the 19th century (Fisher, Neyman, Pearson, etc.)
- First randomized, controlled trial (RCT) by Peirce and Jastrow in 1884
  - First experiment by Gosnell (1924)
  - Gerber and Green (2000) first major *field* experiment

Survey-Experiments

- Rise of surveys in the behavioral revolution
  - Experimentation rare because of paper mode
  - Limited use of "split ballots"

Survey-Experiments

- Rise of surveys in the behavioral revolution
  - Experimentation rare because of paper mode
  - Limited use of "split ballots"
- 1983: Merrill Shanks and the Berkeley Survey Research Center develop **CATI**

Survey-Experiments

- Rise of surveys in the behavioral revolution
  - Experimentation rare because of paper mode
  - Limited use of "split ballots"
- 1983: Merrill Shanks and the Berkeley Survey Research Center develop **CATI**
- Mid-1980s: Paul Sniderman & Tom Piazza performed the first survey experiment[1]
  - Then: the "first multi-investigator"
  - Later: Skip Lupia and Diana Mutz created TESS

[1] Sniderman, Paul M., and Thomas Piazza. 1993. *The Scar of Race*. Cambridge, MA: Harvard University Press.

Survey-experiments, specifically

Survey-experiments, specifically

- A survey experiment is just an experiment that occurs in a survey context
  - As opposed to in the field or in a laboratory

Survey-experiments, specifically

- A survey experiment is just an experiment that occurs in a survey context
  - As opposed to in the field or in a laboratory
- Properties:
  - Sample is representative of population in every respect (in expectation)
  - Sample Average Treatment Effect (SATE) is the average of the sample's individual-level treatment effects
  - SATE is unbiased estimate of PATE

Survey-experiments, specifically

- A survey experiment is just an experiment that occurs in a survey context
  - As opposed to in the field or in a laboratory
- Properties:
  - Sample is representative of population in every respect (in expectation)
  - Sample Average Treatment Effect (SATE) is the average of the sample's individual-level treatment effects
  - SATE is unbiased estimate of PATE
- Sometimes a distinction is made between survey and online experiments

TESS

- Time-Sharing Experiments for the Social Sciences
- Multi-disciplinary initiative that provides infrastructure for survey experiments on nationally representative samples of the United States population
- Funded by the U.S. National Science Foundation
- Anyone anywhere in the world can apply[2]

[2] See also: LISS, Bergen's Citizen Panel, Gothenburg's Citizen Panel

TESS has "Open Protocols"

Protocol is the complete planning document for how to design, implement, and analyze an experiment.[3]

1. Theory/hypotheses
2. Instrumentation
   - Manipulation(s)
   - Outcome(s)
   - Covariate(s)
   - Manipulation check(s)
3. Sampling
4. Implementation
5. Analysis

[3] Thomas J. Leeper. 2011. "The Use of Protocol in the Design and Reporting of Experiments." *The Experimental Political Scientist*.

Why bother writing a protocol?

Why bother writing a protocol?

- Be clear to yourself what you're trying to do before you do it

Why bother writing a protocol?

- Be clear to yourself what you're trying to do before you do it
- Assess the literature for best practices

Why bother writing a protocol?

- Be clear to yourself what you're trying to do before you do it
- Assess the literature for best practices
- Highlight areas in need of pilot testing

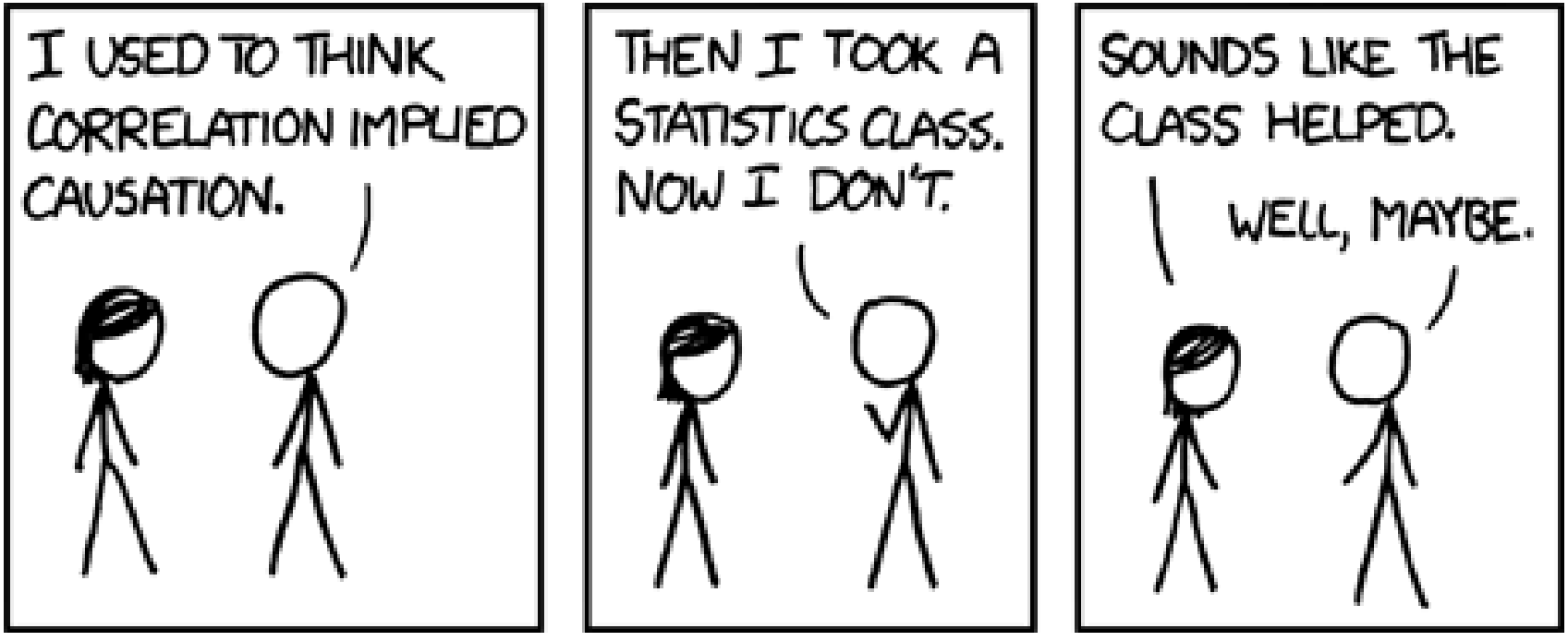
Why bother writing a protocol?

- Be clear to yourself what you're trying to do before you do it
- Assess the literature for best practices
- Highlight areas in need of pilot testing
- Economize questionnaire development

Why bother writing a protocol?

- Be clear to yourself what you're trying to do before you do it
- Assess the literature for best practices
- Highlight areas in need of pilot testing
- Economize questionnaire development
- Study preregistration

Questions?


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

Addressing Confounding

In observational research...

Addressing Confounding

In observational research...

1. Correlate a "putative" cause (X) and an outcome (Y)

Addressing Confounding

In observational research...

1. Correlate a "putative" cause (X) and an outcome (Y)
2. Identify all possible confounds ($\mathbf{Z}$)

Addressing Confounding

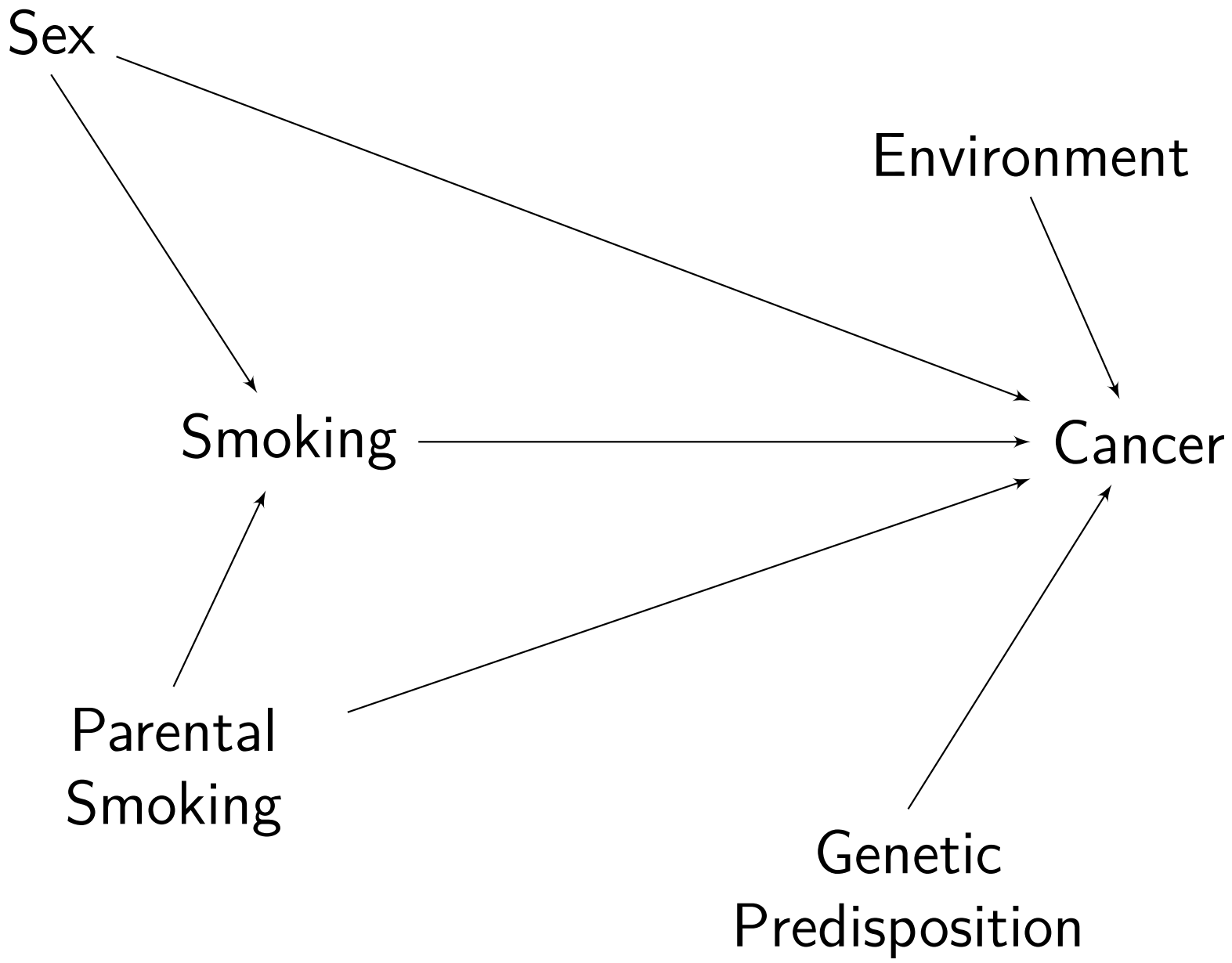
In observational research...

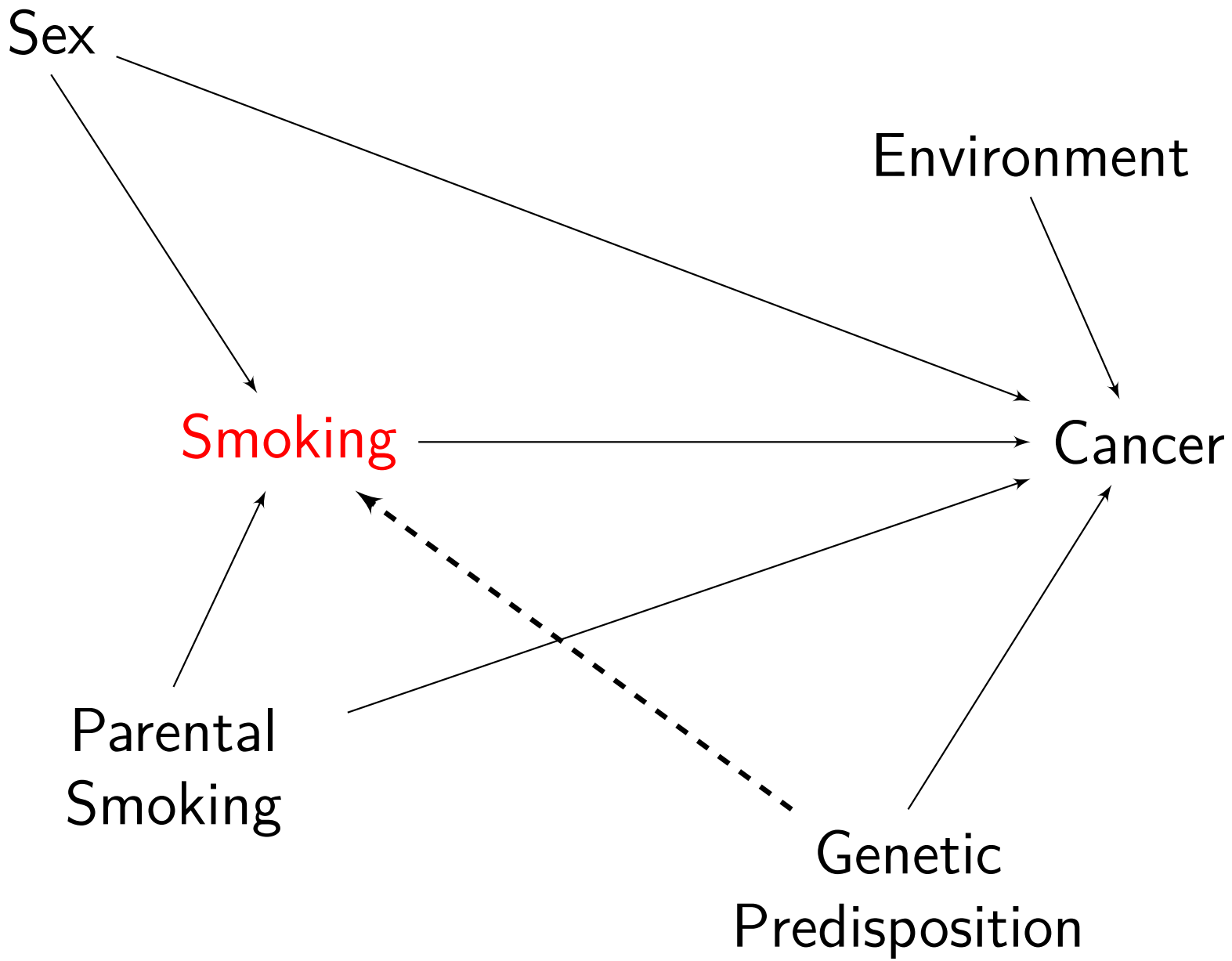
1. Correlate a "putative" cause (X) and an outcome (Y)
2. Identify all possible confounds ($\mathbf{Z}$)
3. "Condition" on all confounds
   - Calculate correlation between X and Y at each combination of levels of $\mathbf{Z}$

Addressing Confounding

In observational research...

1. Correlate a "putative" cause (X) and an outcome (Y)
2. Identify all possible confounds ($\mathbf{Z}$)
3. "Condition" on all confounds
   - Calculate correlation between X and Y at each combination of levels of $\mathbf{Z}$
4. Basically: $Y = \beta_0 + \beta_1 X + \beta Z + \epsilon$


Sex
Environment
Smoking
Cancer
Parental
Smoking
Genetic
Predisposition


Sex
Environment
Smoking
Cancer
Parental
Smoking
Genetic
Predisposition

Experiments are different

Experiments are different

1. Draw causal inferences through *design* not *analysis*

Experiments are different

1. Draw causal inferences through *design* not *analysis*
2. Randomization breaks selection bias

Experiments are different

1. Draw causal inferences through *design* not *analysis*
2. Randomization breaks selection bias
3. We don't need to "control" for anything

Experiments are different

1. Draw causal inferences through *design* not *analysis*
2. Randomization breaks selection bias
3. We don't need to "control" for anything
4. We see "causal effects" in the comparison of experimental groups

Mill's Method of Difference

If an instance in which the phenomenon under investigation occurs, and an instance in which it does not occur, have every circumstance save one in common, that one occurring only in the former; the circumstance in which alone the two instances differ, is the effect, or cause, or an necessary part of the cause, of the phenomenon.

Mill's Method of Difference

If an instance in which the phenomenon under investigation occurs, and an instance in which it does not occur, **have every circumstance save one in common**, that one occurring only in the former; **the circumstance in which alone the two instances differ, is the** effect, or **cause**, or an necessary part of the cause, **of the phenomenon**.

Definitions

Definitions

Unit: A physical object at a particular point in time

Definitions

Treatment: An intervention, whose effect(s) we wish to assess relative to some other (non-)intervention

Definitions

Potential outcomes: The outcome for each unit that we would observe if that unit received each treatment

- Multiple potential outcomes for each unit, but we only observe one of them

Definitions

Causal effect: The comparisons between the unit-level potential outcomes under each intervention

The Experimental Ideal

A randomized experiment, or randomized control trial is:

> *The observation of units after, and possibly before, a randomly assigned intervention in a controlled setting, which tests one or more precise causal expectations*

This is Holland's "statistical solution" to the fundamental problem of causal inference

The Experimental Ideal

- It solves both the temporal ordering and confounding problems of observational causal inference
  - Treatment (X) is applied by the researcher before outcome (Y)
  - Randomization means there are no confounding (Z) variables

The Experimental Ideal

- It solves both the temporal ordering and confounding problems of observational causal inference
  - Treatment (X) is applied by the researcher before outcome (Y)
  - Randomization means there are no confounding (Z) variables
- Thus experiments are a "gold standard" of causal inference

The Experimental Ideal

- It solves both the temporal ordering and confounding problems of observational causal inference
  - Treatment (X) is applied by the researcher before outcome (Y)
  - Randomization means there are no confounding (Z) variables
- Thus experiments are a "gold standard" of causal inference
- Basically: $Y = \beta_0 + \beta_1 X + \epsilon$

Neyman-Rubin Potential Outcomes Framework

If we are interested in some outcome Y, then for every unit i, there are numerous "potential outcomes" $Y*$ only one of which is visible in a given reality. Comparisons of (partially unobservable) potential outcomes indicate causality.

Neyman-Rubin Potential Outcomes Framework

Concisely, we typically discuss two potential outcomes:

- Y_{0i}, the *potential* outcome *realized* if $X_i = 0$ (b/c $D_i = 0$, assigned to control)
- Y_{1i}, the *potential* outcome *realized* if $X_i = 1$ (b/c $D_i = 1$, assigned to treatment)

Historical Aside

- The history of the potential outcomes framework is contested
- Most people attribute it to Donald Rubin
- Paul Holland was the first to link to the philosophical discussions of causality
- Donald Rubin attributes this to Jerzy Neyman (1923)
- James Heckman denies all of this and attributes it Andrew Roy (1951)

Experimental Inference I

- Each unit has multiple *potential* outcomes, but we only observe one of them, randomly

Experimental Inference I

- Each unit has multiple *potential* outcomes, but we only observe one of them, randomly
- In this sense, we are sampling potential outcomes from each unit's population of potential outcomes

unit	low	high
1	?	?
2	?	?
3	?	?
4	?	?

Experimental Inference I

- Each unit has multiple *potential* outcomes, but we only observe one of them, randomly
- In this sense, we are sampling potential outcomes from each unit's population of potential outcomes

unit	low	high	control
1	?	?	?
2	?	?	?
3	?	?	?
4	?	?	?

Experimental Inference I

- Each unit has multiple *potential* outcomes, but we only observe one of them, randomly
- In this sense, we are sampling potential outcomes from each unit's population of potential outcomes

unit	low	high	control	etc.
1	?	?	?	...
2	?	?	?	...
3	?	?	?	...
4	?	?	?	...

Experimental Inference II

- We cannot see individual-level causal effects

Experimental Inference II

- We cannot see individual-level causal effects
- We can see *average causal effects*
  - Ex.: Average difference in cancer between those who do and do not smoke

Experimental Inference II

- We cannot see individual-level causal effects
- We can see *average causal effects*
  - Ex.: Average difference in cancer between those who do and do not smoke
- We want to know: $TE_i = Y_{1i} - Y_{0i}$

Experimental Inference III

- We want to know: $TE_i = Y_{1i} - Y_{0i}$ for every i in the population

Experimental Inference III

- We want to know: $TE_i = Y_{1i} - Y_{0i}$ for every i in the population
- We can average:
  $E[TE_i] = E[Y_{1i} - Y_{0i}] = E[Y_{1i}] - E[Y_{0i}]$

Experimental Inference III

- We want to know: $TE_i = Y_{1i} - Y_{0i}$ for every i in the population
- We can average:
  $E[TE_i] = E[Y_{1i} - Y_{0i}] = E[Y_{1i}] - E[Y_{0i}]$
- But we still only see one potential outcome for each unit:

  $ATE_{naive} = E[Y_{1i}|X = 1] - E[Y_{0i}|X = 0]$

Experimental Inference III

- We want to know: $TE_i = Y_{1i} - Y_{0i}$ for every i in the population
- We can average:
  $E[TE_i] = E[Y_{1i} - Y_{0i}] = E[Y_{1i}] - E[Y_{0i}]$
- But we still only see one potential outcome for each unit:

  $ATE_{naive} = E[Y_{1i}|X = 1] - E[Y_{0i}|X = 0]$
- Is this what we want to know?

Experimental Inference IV

- What we want and what we have:

$$ATE = E[Y_{1i}] - E[Y_{0i}] \tag{1}$$

$$ATE_{naive} = E[Y_{1i}|X = 1] - E[Y_{0i}|X = 0] \tag{2}$$

Experimental Inference IV

- What we want and what we have:

$$ATE = E[Y_{1i}] - E[Y_{0i}] \tag{1}$$

$$ATE_{naive} = E[Y_{1i}|X = 1] - E[Y_{0i}|X = 0] \tag{2}$$

- Are the following statements true?
  - $E[Y_{1i}] = E[Y_{1i}|X = 1]$
  - $E[Y_{0i}] = E[Y_{0i}|X = 0]$

Experimental Inference IV

- What we want and what we have:

$$ATE = E[Y_{1i}] - E[Y_{0i}] \tag{1}$$

$$ATE_{naive} = E[Y_{1i}|X = 1] - E[Y_{0i}|X = 0] \tag{2}$$

- Are the following statements true?
  - $E[Y_{1i}] = E[Y_{1i}|X = 1]$
  - $E[Y_{0i}] = E[Y_{0i}|X = 0]$
- Not in general!

Experimental Inference V

- Only true when both of the following hold:

$$E[Y_{1i}] = E[Y_{1i}|X = 1] = E[Y_{1i}|X = 0] \quad (3)$$
$$E[Y_{0i}] = E[Y_{0i}|X = 1] = E[Y_{0i}|X = 0] \quad (4)$$

- In that case, potential outcomes are *independent* of treatment assignment
- If true (e.g., due to randomization of X), then:

$$\begin{aligned} ATE_{naive} &= E[Y_{1i}|X = 1] - E[Y_{0i}|X = 0] \quad (5) \\ &= E[Y_{1i}] - E[Y_{0i}] \\ &= ATE \end{aligned}$$

Experimental Inference VI

- This holds in experiments because of a *physical process of randomization*[4]

[4] Random means "known probability of treatment" not "haphazard".

Experimental Inference VI

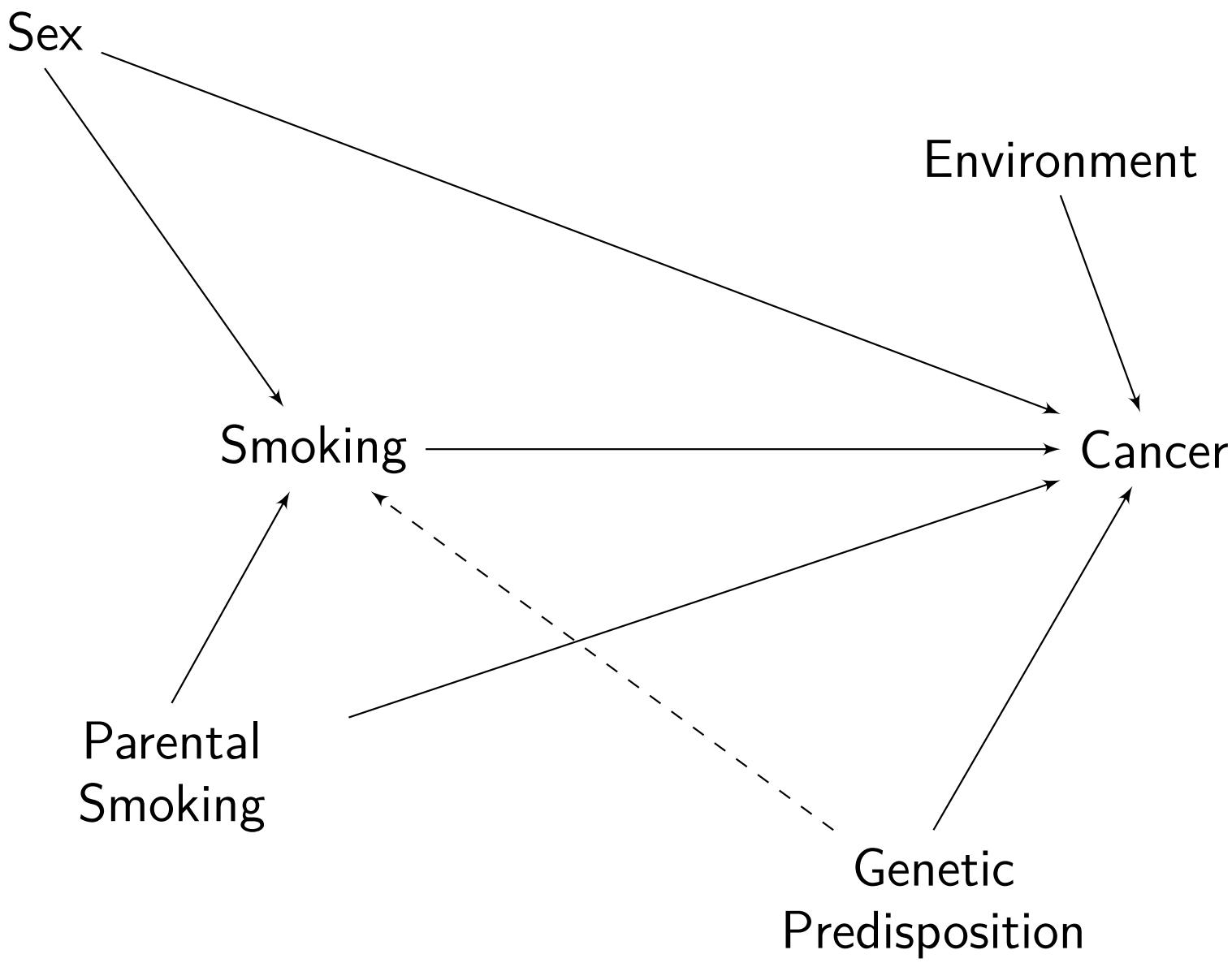
- This holds in experiments because of a *physical process of randomization*[4]
- Units differ only in side of coin that was up
  - $X_i = 1$ only because $D_i = 1$

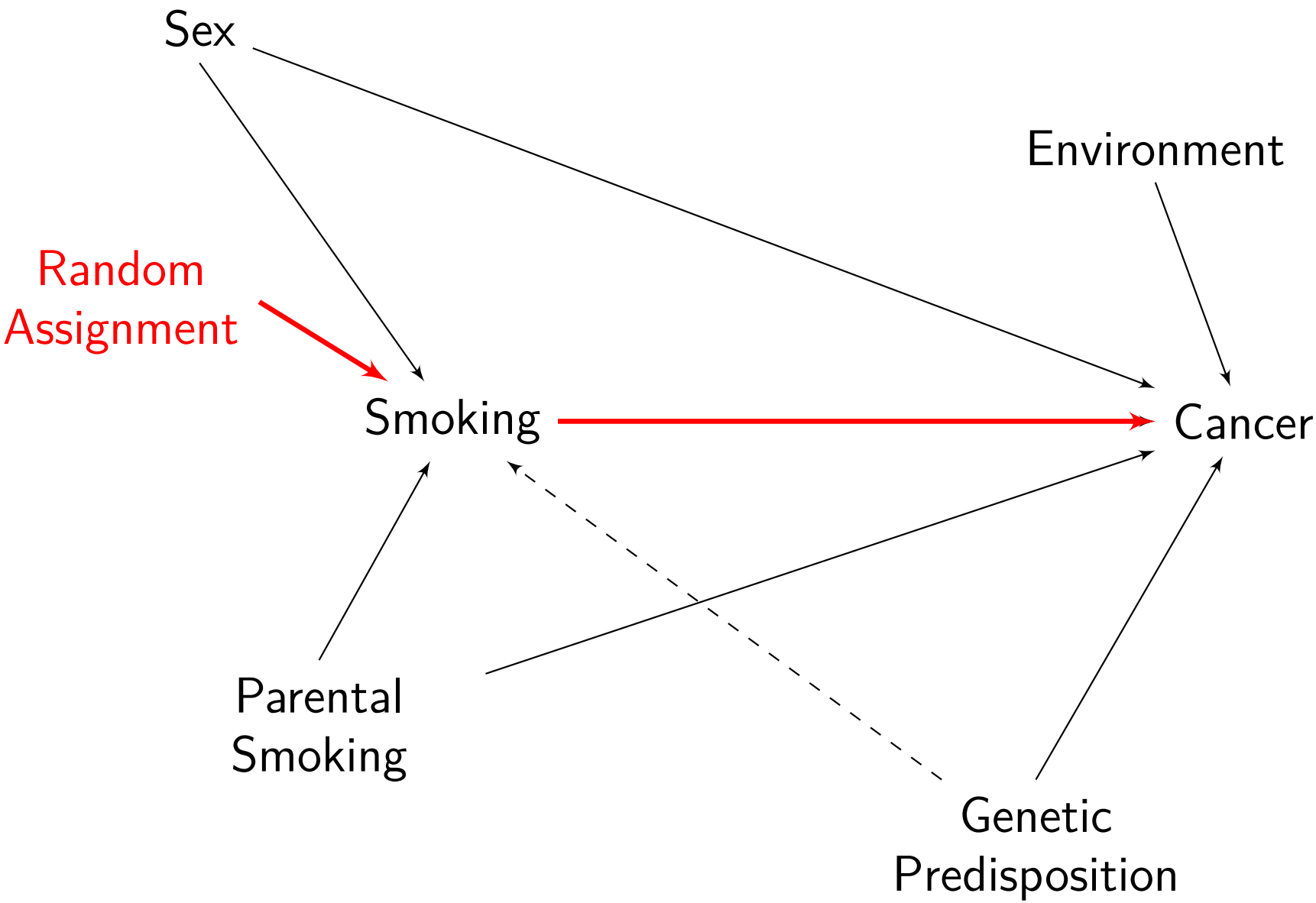
[4] Random means "known probability of treatment" not "haphazard".

Experimental Inference VI

- This holds in experiments because of a *physical process of randomization*[4]
- Units differ only in side of coin that was up
  - $X_i = 1$ only because $D_i = 1$
- Implications:
  - Covariate balance
  - Potential outcomes balanced and independent of treatment assignment
  - No confounding (selection bias)

[4] Random means "known probability of treatment" not "haphazard".


Sex
Environment
Smoking
Cancer
Parental
Smoking
Genetic
Predisposition


Sex
Environment
Random
Assignment
Smoking
Cancer
Parental
Smoking
Genetic
Predisposition

Questions?

Experimental Analysis I

- The statistic of interest in an experiment is the *sample average treatment effect* (SATE)
- If our sample is *representative*, then this provides an estimate of the population average treatment (PATE)
- This boils down to being a mean-difference between two groups:

$$SATE = \frac{1}{n_1} \sum Y_{1i} - \frac{1}{n_0} \sum Y_{0i} \tag{5}$$

Computation of Effects I

- In practice we often estimate SATE using t-tests, ANOVA, or OLS regression
- These are all basically equivalent

Computation of Effects I

- In practice we often estimate SATE using t-tests, ANOVA, or OLS regression
- These are all basically equivalent
- Reasons to choose one procedure over another:
  - Disciplinary norms

Computation of Effects I

- In practice we often estimate SATE using t-tests, ANOVA, or OLS regression
- These are all basically equivalent
- Reasons to choose one procedure over another:
  - Disciplinary norms
  - Ease of interpretation

Computation of Effects I

- In practice we often estimate SATE using t-tests, ANOVA, or OLS regression
- These are all basically equivalent
- Reasons to choose one procedure over another:
  - Disciplinary norms
  - Ease of interpretation
  - Flexibility for >2 treatment conditions

Computation of Effects II

An experimental data structure looks like:

unit	treatment	outcome
1	0	13
2	0	6
3	0	4
4	0	5
5	1	3
6	1	1
7	1	10
8	1	9

Computation of Effects II

Sometimes it looks like this instead, which is bad:

unit	treatment	outcome0	outcome1
1	0	13	NA
2	0	6	NA
3	0	4	NA
4	0	5	NA
5	1	NA	3
6	1	NA	1
7	1	NA	10
8	1	NA	9

Computation of Effects II

An experimental data structure looks like:

unit	treatment	outcome
1	0	13
2	0	6
3	0	4
4	0	5
5	1	3
6	1	1
7	1	10
8	1	9

Computation of Effects III

R:

```
t.test(outcome ~ treatment, data = data)
lm(outcome ~ factor(treatment), data = data)
```

Stata:

```
ttest outcome, by(treatment)
reg outcome i.treatment
```

Questions?

Let's work in R!

(Basic analysis)

Experimental Analysis II

- We don't just care about the size of the SATE. We also want to know whether it is significantly different from zero (i.e., different from no effect/difference)
- To know that, we need to estimate the *variance* of the SATE
- The variance is influenced by:
  - Total sample size
  - Variance of the outcome, Y
  - Relative size of each treatment group

Experimental Analysis III

- Formula for the variance of the SATE is:
$$\widehat{Var}(SATE) = \frac{\widehat{Var}(Y_0)}{n_0} + \frac{\widehat{Var}(Y_1)}{n_1}$$
  - $\widehat{Var}(Y_0)$ is control group variance
  - $\widehat{Var}(Y_1)$ is treatment group variance
- We often express this as the *standard error* of the estimate:
$$\widehat{SE}_{SATE} = \sqrt{\frac{\widehat{Var}(Y_0)}{n_0} + \frac{\widehat{Var}(Y_1)}{n_1}}$$

Intuition about Variance

- Bigger sample $\rightarrow$ smaller SEs
- Smaller variance $\rightarrow$ smaller SEs
- Efficient use of sample size:
  - When treatment group variances equal, equal sample sizes are most efficient
  - When variances differ, sample units are better allocated to the group with higher variance in Y

Statistical Power

- Power analysis to determine sample size
- Type I and Type II Errors
  - True positive rate is power
  - False negative rate is the significance threshold (α)

	H_0 True	H_0 False
Reject H_0	Type 1 Error	**True positive**
Accept H_0	False negative	Type II error

Doing a Power Analysis

- μ, Treatment group mean outcomes
- N, Sample size
- σ, Outcome variance
- α Statistical significance threshold
- ϕ, a sampling distribution

$$Power = \phi\left(\frac{|\mu_1 - \mu_0|\sqrt{N}}{2\sigma} - \phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right)$$

Intuition about Power

Minimum detectable effect is the smallest effect we could detect given sample size, "true" effect size, variance of outcome, power, and α.

In essence: some non-zero effect sizes are not detectable by a study of a given sample size.[5]

[5] Gelman, A. and Weakliem, D. 2009. "Of Beauty, Sex and Power." *American Scientist* 97(4): 310–16

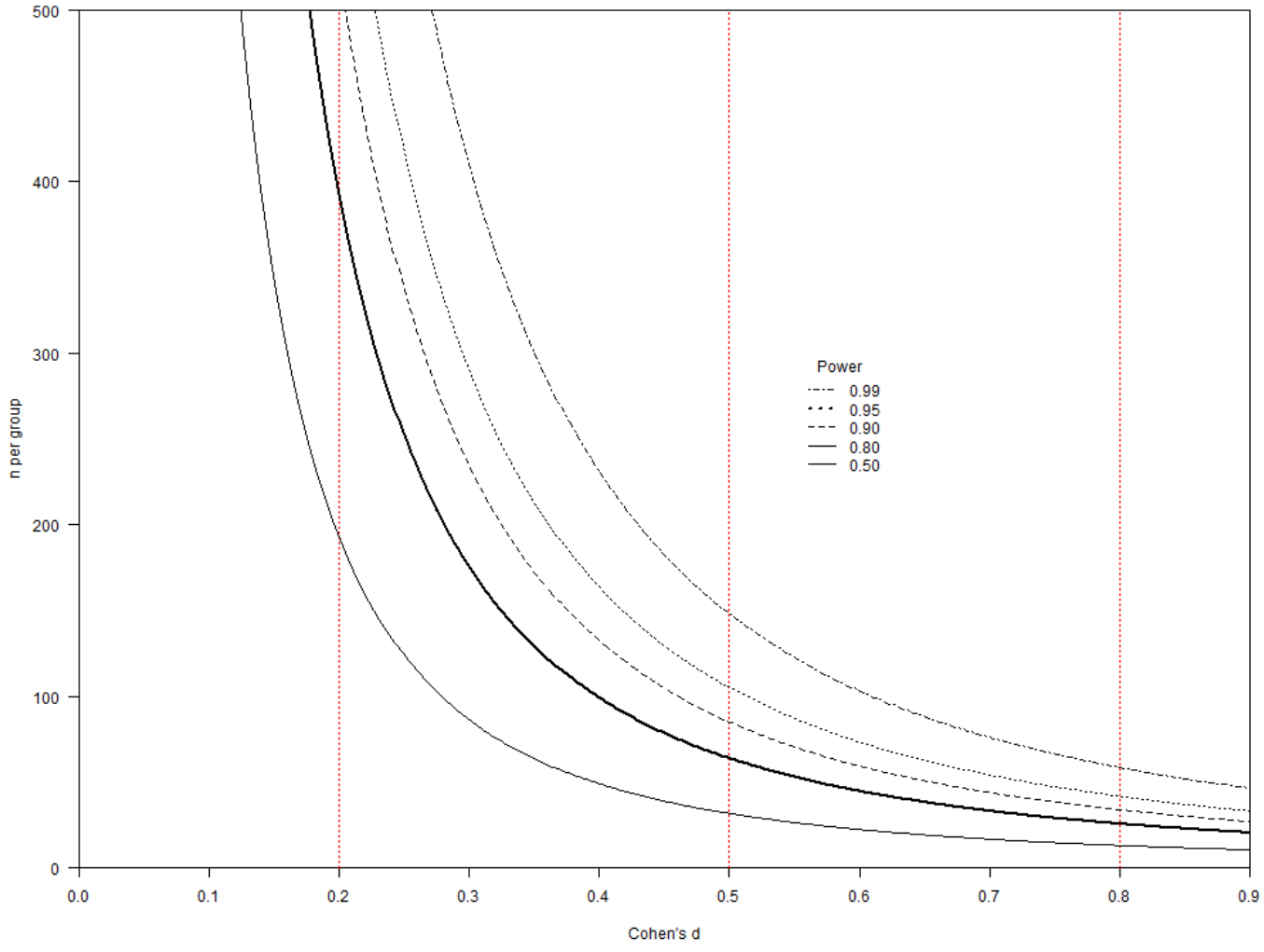
Intuition about Power

- It can help to think in terms of "standardized effect sizes"
- Cohen's d:
  $d = \frac{\bar{x}_1 - \bar{x}_0}{s}$, where $s = \sqrt{\frac{(n_1-1)s_1^2+(n_0-1)s_0^2}{n_1+n_0-2}}$
- Intuition: How large is the effect in standard deviations of the outcome?
  - Know if effects are large or small
  - Compare effects across studies
- Small: 0.2; Medium: 0.5; Large: 0.8

Let's work in R!

(Power Analysis)

Intuition about Power



What kinds of questions can we answer with (survey) experiments?

What kinds of questions can we answer with (survey) experiments?

- Forward causal questions
  - Can X cause Y?
  - What effects does X have?

What kinds of questions can we answer with (survey) experiments?

- Forward causal questions
  - Can X cause Y?
  - What effects does X have?
- Backward causal questions
  - What causes Y?
  - How much of Y is attributable to X?

What kinds of questions can we answer with (survey) experiments?

- Forward causal questions
  - Can X cause Y?
  - What effects does X have?
- Backward causal questions
  - What causes Y?
  - How much of Y is attributable to X?
- Even though answering "forward" causal question, we start with an outcome concept

Hypothesis Testing

- From theory, we derive testable hypotheses
  - Hypotheses are expectations about differences in outcomes across levels of a putatively causal variable
  - Hypothesis must be testable by an SATE
- Manipulations are developed to create variation in that causal variable

Example: News Framing

- Theory: Presentation of news affects opinion
- Hypotheses:
  - News emphasizing free speech increases support for a hate group rally
  - News emphasizing public safety decreases support for a hate group rally
- Manipulation:
  - Control group: no information
  - Free speech group: article emphasizing rights
  - Public safety group: article emphasizing safety

Example: Partisan Identity

- Theory: Strength of partisan identity affects tendency to accept party position
- Hypotheses:
  - Strong partisans are more likely to accept their party's position on an issue
- Manipulation:
  - Control group: no manipulation
  - "Univalent" condition
  - "Ambivalent" condition

Univalent

These days, Democrats and Republicans differ from one another considerably. The two groups seem to be growing further and further apart, not only in terms of their opinions but also their lifestyles. Earlier in the survey, you said you tend to identify as a *Democrat/ Republican*. Please take a few minutes to think about what you like about *Democrats/ Republicans* compared to the *Republicans/ Democrats*. Think of 2 to 3 things you especially like best about **your party**. Then think of 2 to 3 things you especially dislike about **the other party**. Now please write those thoughts in the space below.

Ambivalent

These days, Democrats and Republicans differ from one another considerably. The two groups seem to be growing further and further apart, not only in terms of their opinions but also their lifestyles. Earlier in the survey, you said you tend to identify as a *Democrat/ Republican*. Please take a few minutes to think about what you like about *Democrats/ Republicans* compared to the *Republicans/ Democrats*. Think of 2 to 3 things you especially like best about **the other party**. Then think of 2 to 3 things you especially dislike about **your party**. Now please write those thoughts in the space below.

Treatments Test Hypotheses!

Treatments Test Hypotheses!

- Derive experimental design from hypotheses

Treatments Test Hypotheses!

- Derive experimental design from hypotheses
- Experimental “factors” are expressions of hypotheses as randomized groups

Treatments Test Hypotheses!

- Derive experimental design from hypotheses
- Experimental "factors" are expressions of hypotheses as randomized groups
- What intervention each group receives depends on hypotheses
  - presence/absence
  - levels/doses
  - qualitative variations

Ex.: Presence/Absence

- Theory: Negative campaigning reduces support for the party described negatively.
- Hypothesis: Exposure to a negative advertisement criticizing a party reduces support for that party.
- Manipulation:
  - Control group receives no advertisement.
  - Treatment group watches a video containing a negative ad describing a party.

Ex.: Levels/doses

- Theory: Negative campaigning reduces support for the party described negatively.
- Hypothesis: Exposure to higher levels of negative advertising criticizing a party reduces support for that party.
- Manipulation:
  - Control group receives no advertisement.
  - Treatment group 1 watches a video containing 1 negative ad describing a party.
  - Treatment group 2 watches a video containing 2 negative ads describing a party.
  - Treatment group 3 watches a video containing 3 negative ads describing a party.
  - etc.

Ex.: Qualitative variation

- Theory: Negative campaigning reduces support for the party described negatively.
- Hypothesis: Exposure to a negative advertisement criticizing a party reduces support for that party, while a positive advertisement has no effect.
- Manipulation:
  - Control group receives no advertisement.
  - Negative treatment group watches a video containing a negative ad describing a party.
  - Positive treatment group watches a video containing a positive ad describing a party.

Questions?

Activity!

- How do we know if an experiment is any good?
- Talk with a partner for about 3 minutes
- Try to develop some criteria that allow you to evaluate “what makes for a good experiment?”

Some possible criteria

- Significant results
- Face validity
- Coherent for respondents
- Non-obvious to respondents
- Simple
- Indirect/unobtrusive
- Validated by prior work
- Innovative/creative
- ...

The best criterion for evaluating the quality of an experiment is whether it manipulated the intended independent variable and controlled everything else by design.

The best criterion for evaluating the quality of an experiment is whether it manipulated the intended independent variable and controlled everything else by design.

–Thomas J. Leeper (18 January 2017)

How do we know we manipulated what we think we manipulated?

How do we know we manipulated what we think we manipulated?

- Outcomes are affected consistent with theory

How do we know we manipulated what we think we manipulated?

- Outcomes are affected consistent with theory
- Before the study using *pilot testing* (or *pretesting*)

How do we know we manipulated what we think we manipulated?

- Outcomes are affected consistent with theory
- Before the study using *pilot testing* (or *pretesting*)
- During the study, using *manipulation checks*

How do we know we manipulated what we think we manipulated?

- Outcomes are affected consistent with theory
- Before the study using *pilot testing* (or *pretesting*)
- During the study, using *manipulation checks*
- During the study, using *placebos*

How do we know we manipulated what we think we manipulated?

- Outcomes are affected consistent with theory
- Before the study using *pilot testing* (or *pretesting*)
- During the study, using *manipulation checks*
- During the study, using *placebos*
- During the study, using *non-equivalent outcomes*

I. Outcomes Affected

- Follows a circular logic!
- Doesn't tell us anything if we hypothesize null effects

II. Pilot Testing

- Goal: establish construct validity of manipulation
- Assess whether a set of possible manipulations affect a measure of the *independent* variable

II. Pilot Testing

- Goal: establish construct validity of manipulation
- Assess whether a set of possible manipulations affect a measure of the *independent* variable
- Example:
  - Goal: Manipulate the "strength" of an argument
  - Write several arguments
  - Ask pilot test respondents to report how strong each one was

III. Manipulation Checks

- Manipulation checks are items added post-treatment, post-outcome that assess whether the *independent* variable was affected by treatment
- We typically talk about manipulations as directly setting the value of X, but in practice we are typically manipulating something *that we think* strongly modifies X

III. Manipulation Checks

- Manipulation checks are items added post-treatment, post-outcome that assess whether the *independent* variable was affected by treatment
- We typically talk about manipulations as directly setting the value of X, but in practice we are typically manipulating something *that we think* strongly modifies X
- Example: information manipulations aim to modify knowledge or beliefs, but are necessarily imperfect at doing so

Manipulation check example[6]

1. Treatment 1: Supply Information
2. Manipulation check 1: measure beliefs
3. Treatment 2: Prime a set of considerations
4. Outcome: Measure opinion
5. Manipulation check 2: measure dimension salience

[6] Leeper & Slothuus. n.d. "Can Citizens Be Framed?" Available from: `http://thomasleeper.com/research.html`.

Some Best Practices

Some Best Practices

- Manipulation checks should be innocuous
  - Shouldn't modify independent variable
  - Shouldn't modify outcome variable

Some Best Practices

- Manipulation checks should be innocuous
  - Shouldn't modify independent variable
  - Shouldn't modify outcome variable
- Generally, measure post-outcome

Some Best Practices

- Manipulation checks should be innocuous
  - Shouldn't modify independent variable
  - Shouldn't modify outcome variable
- Generally, measure post-outcome
- Measure both what you wanted to manipulate *and* what you didn't want to manipulate
  - Most treatments are *compound*!

IV. Placebos

- Include an experimental condition that *does not* manipulate the variable of interest (but might affect the outcome)

IV. Placebos

- Include an experimental condition that *does not* manipulate the variable of interest (but might affect the outcome)
- Example:
  - Study whether risk-related arguments about climate change increase support for a climate change policy
  - Placebo condition: control article with risk-related arguments about non-environmental issue (e.g., terrorism)

V. Non-equivalent outcomes

- Measures an outcome that *should not* be affected by independent variable

V. Non-equivalent outcomes

- Measures an outcome that *should not* be affected by independent variable
- Example:
  - Assess effect of some treatment on attitudes toward group A
  - Focal outcome: attitudes toward group A
  - Non-equivalent outcome: attitudes toward group B

Aside: Demand Characteristics

- "Demand characteristics" are features of experiments that (unintentionally) imply the purpose of the study and thereby change respondents' behavior (to be consistent with theory)

[7] But, consider the ethics of not doing so (more Friday)

Aside: Demand Characteristics

- "Demand characteristics" are features of experiments that (unintentionally) imply the purpose of the study and thereby change respondents' behavior (to be consistent with theory)
- Implications:
  - Design experimental treatments that are non-obvious
  - Do not disclose the purpose of the study up front[7]

[7] But, consider the ethics of not doing so (more Friday)

Question Wording Designs

- Kahneman and Tversky used a lot of "question wording" experiments
- Hypothesized difference in outcomes according to the decision being faced
  - Risky or not risky
  - Gains or losses
- Manipulation operationalizes this by asking two different questions
- Outcome is the answer to the question

"Framing" or "Priming" Experiments

Example: Schuldt et al. "'Global Warming' or 'Climate Change'? Whether the Planet is Warming Depends on Question Wording."

What's this study about?

You may have heard about the idea that the world's temperature may have been **going up** over the past 100 years, a phenomenon sometimes called **global warming**. What is your personal opinion regarding whether or not this has been happening?

- Definitely has not been happening
- Probably has not been happening
- Unsure, but leaning toward it has not been happening
- Not sure either way
- Unsure, but leaning toward it has been happening
- Probably has been happening
- Definitely has been happening

You may have heard about the idea that the world's temperature may have been **changing** over the past 100 years, a phenomenon sometimes called **climate change**. What is your personal opinion regarding whether or not this has been happening?

- Definitely has not been happening
- Probably has not been happening
- Unsure, but leaning toward it has not been happening
- Not sure either way
- Unsure, but leaning toward it has been happening
- Probably has been happening
- Definitely has been happening

Another framing example[8]

Today, tests are being developed that make it possible to detect serious genetic defects **before a baby is born**. But so far, it is impossible either to treat or to correct most of them. If (you/your partner) were pregnant, would you want (her) to have a test to find out if the **baby** has any serious genetic defects? (Yes/No)

Suppose a test shows the **baby** has a serious genetic defect. Would you, yourself, want (your partner) to have an abortion if a test shows the **baby** has a serious genetic defect? (Yes/No)

[8] Singer & Couper. 2014. "The Effect of Question Wording on Attitudes toward Prenatal Testing and Abortion." *Public Opinion Quarterly* 78(3): 751–760.

Another framing example[8]

Today, tests are being developed that make it possible to detect serious genetic defects **in the fetus during pregnancy**. But so far, it is impossible either to treat or to correct most of them. If (you/your partner) were pregnant, would you want (her) to have a test to find out if the **fetus** has any serious genetic defects? (Yes/No)

Suppose a test shows the **fetus** has a serious genetic defect. Would you, yourself, want (your partner) to have an abortion if a test shows the **fetus** has a serious genetic defect? (Yes/No)

[8] Singer & Couper. 2014. "The Effect of Question Wording on Attitudes toward Prenatal Testing and Abortion." *Public Opinion Quarterly* 78(3): 751–760.

Another framing example[9]

Do you favor or oppose the death penalty for persons convicted of murder?

[9] Bobo & Johnson. 2004. "A Taste for Punishment: Black and White Americans' Views on the Death Penalty and the War on Drugs." Du Bois Review 1(1): 151–180.

Another framing example[9]

Blacks are about 12% of the U.S. population, but they were half of the homicide offenders last year. Do you favor or oppose the death penalty for persons convicted of murder?

[9] Bobo & Johnson. 2004. "A Taste for Punishment: Black and White Americans' Views on the Death Penalty and the War on Drugs." Du Bois Review 1(1): 151–180.

Another framing example[10]

Concealed handgun laws have recently received national attention. Some people have argued that law-abiding citizens have the right to protect themselves. What do you think about concealed handgun laws?

[10] Haider-Markel & Joslyn. 2001. "Gun Policy, Opinion, Tragedy, and Blame Attribution: The Conditional Influence of Issue Frames." *Journal of Politics* 63(2): 520–543.

Another framing example[10]

Concealed handgun laws have recently received national attention. Some people have argued that laws allowing citizens to carry concealed handguns threaten public safety because they would allow almost anyone to carry a gun almost anywhere, even onto school grounds. What do you think about concealed handgun laws?

[10] Haider-Markel & Joslyn. 2001. "Gun Policy, Opinion, Tragedy, and Blame Attribution: The Conditional Influence of Issue Frames." *Journal of Politics* 63(2): 520–543.

Question testing

Use question wording designs to select which survey measures we want to use

- Select possible question wordings
- Select some criterion(-ia) for assessing which is better
- Pilot test and then use the item that performs better

Aside: Experimentation vs. Other Pretesting Methods

Aside: Experimentation vs. Other Pretesting Methods

- Experiments are complementary to other pretesting methods

Aside: Experimentation vs. Other Pretesting Methods

- Experiments are complementary to other pretesting methods
- Specific value added of an experiment: optimize questions or other survey features against a specific criterion, e.g.:
  - (Non-)Response or drop-off rates
  - "Don't know" rates
  - Item characteristics
  - Reading times or response latencies

Aside: Experimentation vs. Other Pretesting Methods

- Experiments are complementary to other pretesting methods
- Specific value added of an experiment: optimize questions or other survey features against a specific criterion, e.g.:
  - (Non-)Response or drop-off rates
  - "Don't know" rates
  - Item characteristics
  - Reading times or response latencies
- But! Power considerations...

Classic question testing experiment[11]

Some people feel that The 1975 Public Affairs Act should be repealed-do you agree or disagree with this idea?

[11] Bishop, G.F., Tuchfarber, A. & Oldendick, R.W. 1986. "Opinions on Fictitious Issues: The Pressure to Answer Survey Questions." *Public Opinion Quarterly* 50(2): 240–250.

Classic question testing experiment[11]

Some people feel that The 1975 Public Affairs Act should be repealed-do you agree or disagree with this idea, or haven't you thought much about this issue?

[11] Bishop, G.F., Tuchfarber, A. & Oldendick, R.W. 1986. "Opinions on Fictitious Issues: The Pressure to Answer Survey Questions." *Public Opinion Quarterly* 50(2): 240–250.

An example[12]

In talking to people about elections, we often find that a lot of people were not able to vote because they weren't registered, they were sick, or they just didn't have time. How about you–did you vote in the elections this November?

[12] Holbrook & Krosnick. 2013. "A New Question Sequence to Measure Voter Turnout in Telephone Surveys: Results of an Experiment in the 2006 ANES Pilot Study." *Public Opinion Quarterly* 77: 106–123.

An example[12]

In talking to people about elections, we often find that a lot of people were not able to vote because they weren't registered, they were sick, or they just didn't have time. Which of the following statements best describes you?

- One, I did not vote in the November 3 election
- two, I thought about voting this time but didn't
- three, I usually vote but didn't this time
- four, I am sure I voted

[12] Holbrook & Krosnick. 2013. "A New Question Sequence to Measure Voter Turnout in Telephone Surveys: Results of an Experiment in the 2006 ANES Pilot Study." *Public Opinion Quarterly* 77: 106–123.

An Instructional Manipulation[13]

For the next few questions, I am going to read out some statements, and for each one, please tell me if it is true or false. If you don't know, just say so and we will skip to the next one.

1. Britain's electoral system is based on proportional representation.
2. MPs from different parties are on parliamentary committees.
3. The Conservatives are opposed to the ratification of a constitution for the European Union.

[13] Sturgis, Allum & Smith. 2008. "An Experiment on the Measurement of Political Knowledge in Surveys." *Public Opinion Quarterly* 72(1): 90–102.

An Instructional Manipulation[13]

For the next few questions, I am going to read out some statements, and for each one, please tell me if it is true or false. If you don't know, please just give me your best guess.

1. Britain's electoral system is based on proportional representation.
2. MPs from different parties are on parliamentary committees.
3. The Conservatives are opposed to the ratification of a constitution for the European Union.

[13] Sturgis, Allum & Smith. 2008. "An Experiment on the Measurement of Political Knowledge in Surveys." *Public Opinion Quarterly* 72(1): 90–102.

An Instructional Manipulation + [14]

In the next part of this study, you will be asked 14 questions about politics, public policy, and economics. Many people don't know the answers to these questions, but it is helpful for us if you answer, even if you're not sure what the correct answer is. We encourage you to take a guess on every question. At the end of this study, you will see a summary of how many questions you answered correctly.

[14] Prior & Lupia. 2008. "Money, Time, and Political Knowledge: Distinguishing Quick Recall and Political Learning Skills." *American journal of Political Science* 52(1): 169–183.

An Instructional Manipulation + [14]

We will pay you for answering questions correctly. You will earn $1 for every correct answer you give. So, if you answer 3 of the 14 questions correctly, you will earn $3. If you answer 7 of the 14 questions correctly, you will earn $7. The more questions you answer correctly, the more you will earn.

[14] Prior & Lupia. 2008. "Money, Time, and Political Knowledge: Distinguishing Quick Recall and Political Learning Skills." *American journal of Political Science* 52(1): 169–183.

Question Order Designs

- Manipulation of pre-outcome questionnaire

Question Order Designs

- Manipulation of pre-outcome questionnaire
- Example:
  - Goal: assess influence of value salience on support for a policy
  - Manipulate by asking different questions:
    - Battery of 5 "rights" questions, or
    - Battery of 5 "life" questions
  - Measure support for legalized abortion

Question Order Designs

- Manipulation of pre-outcome questionnaire
- Example:
  - Goal: assess influence of value salience on support for a policy
  - Manipulate by asking different questions:
    - Battery of 5 "rights" questions, or
    - Battery of 5 "life" questions
  - Measure support for legalized abortion
- If answers to manipulated questions matter, can measure rest post-outcome

Ex. Question-as-treatment[15]

- How close do you feel to your ethnic or racial group?
- Some people have said that taxes need to be raised to take care of pressing national needs. How willing would you be to have your taxes raised to improve education in public schools?

[15] Transue. 2007. "Identity Salience, Identity Acceptance, and Racial Policy Attitudes: American National Identity as a Uniting Force." *American Journal of Political Science* 51(1): 78–91.

Ex. Question-as-treatment[15]

- How close do you feel to other Americans?
- Some people have said that taxes need to be raised to take care of pressing national needs. How willing would you be to have your taxes raised to improve education in public schools?

[15] Transue. 2007. "Identity Salience, Identity Acceptance, and Racial Policy Attitudes: American National Identity as a Uniting Force." *American Journal of Political Science* 51(1): 78–91.

Ex. Question-as-treatment[15]

- How close do you feel to your ethnic or racial group?
- Some people have said that taxes need to be raised to take care of pressing national needs. How willing would you be to have your taxes raised to improve educational opportunities for minorities?

[15] Transue. 2007. "Identity Salience, Identity Acceptance, and Racial Policy Attitudes: American National Identity as a Uniting Force." *American Journal of Political Science* 51(1): 78–91.

Ex. Question-as-treatment[15]

- How close do you feel to other Americans?
- Some people have said that taxes need to be raised to take care of pressing national needs. How willing would you be to have your taxes raised to improve educational opportunities for minorities?

[15] Transue. 2007. "Identity Salience, Identity Acceptance, and Racial Policy Attitudes: American National Identity as a Uniting Force." *American Journal of Political Science* 51(1): 78–91.

Ex.: Knowledge and Political Interest

1. Do you happen to remember anything special that your U.S. Representative has done for your district or for the people in your district while he has been in Congress?
2. Is there any legislative bill that has come up in the House of Representatives, on which you remember how your congressman has voted in the last couple of years?
3. Now, some people seem to follow what's going on in government and public affairs most of the time, whether there's an election going on or not. Others aren't that interested. Would you say that you follow what's going on in government and public affairs most of the time, some of the time, only now and then, or hardly at all?

Ex.: Knowledge and Political Interest

1. Now, some people seem to follow what's going on in government and public affairs most of the time, whether there's an election going on or not. Others aren't that interested. Would you say that you follow what's going on in government and public affairs most of the time, some of the time, only now and then, or hardly at all?
2. Do you happen to remember anything special that your U.S. Representative has done for your district or for the people in your district while he has been in Congress?
3. Is there any legislative bill that has come up in the House of Representatives, on which you remember how your congressman has voted in the last couple of years?

Vignettes

- A "vignette" is a short paragraph of text describing a situation
- Vignettes are probably the most common survey experimental paradigm, after question wording designs
- Take many forms and increasingly encompass non-textual stimuli
- Basically limited to web-based mode

A classic vignette[16]

Now think about a **(black/white)** woman in her early thirties. She is a high school **(graduate/drop out)** with a ten-year-old child, and she has been on welfare for the past year.

- How likely is it that she will have more children in order to get a bigger welfare check? (1 = Very likely, . . . , 7 = Not at all likely)
- How likely do you think it is that she will really try hard to find a job in the next year? (1 = Very likely, . . . , 7 = Not at all likely)

[16] Gilens, M. 1996. "'Race coding' and white opposition to welfare. *American Political Science Review* 90(3): 593–604.

Newer vignette[17]

Imagine that you were living in a village in another district in Uttar Pradesh and that you were voting for candidates in **(village/state/national)** election. Here are the two candidates who are running against each other: The first candidate is named **(caste name)** and is running as the **(BJP/SP/BSP)** party candidate. **(Corrupt/criminality allegation)**. His opponent is named **(caste name)** and is running as the **(BJP/SP/BSP)** party candidate. **(Opposite corrupt/criminality allegation)**. From this information, please indicate which candidate you would vote for in the **(village/state/national)** election.

[17] Banerjee et al. 2012. "Are Poor Voters Indifferent to Whether Elected Leaders are Criminal or Corrupt? A Vignette Experiment in Rural India." Working paper.

Longer texts[18]

We are testing materials for use in a study **of the structure of sentences people use when writing news editorials**. Along these lines, we would like you to read a series of paragraphs, taken from recent major newspaper editorials.

[18] Druckman & Leeper. 2012. "Learning More from Political Communication Experiments: Pretreatment and Its Effects." *American Journal of Political Science* 56(4): 875–896.

Longer texts[18]

We are testing materials for use in a study **that is related to the kinds of opinions people form about public policies**. Along these lines, we would like you to read a series of paragraphs, taken from recent major newspaper editorials.

[18] Druckman & Leeper. 2012. "Learning More from Political Communication Experiments: Pretreatment and Its Effects." *American Journal of Political Science* 56(4): 875–896.

Please read the following paragraphs and, for each, rate **how *dynamic* you think it is. A paragraph is more "dynamic" when it uses more vivid action words. For example, a statement like, "He sped up and raced through the light before crashing into the swerving truck," seems more dynamic than, "He went faster to get through the light before having an accident." The action words in the first sentence (which we have highlighted in bold) seem more dynamic or vivid than those contained in the second sentence.** There are no right or wrong opinions and your responses to all questions are completely confidential.

Please read the following paragraphs and, for each, rate **the extent to which it decreases or increases your support for the Patriot Act. In subsequent surveys we will ask you for your overall opinion about the state-run casino (i.e., the extent to which you oppose or support the state-run casino).** There are no right or wrong opinions and your responses to all questions are completely confidential.

Please read the paragraphs carefully and, after each one, rate **the extent to which you think it is *dynamic***.

With the passage of the Patriot Act in 2001, the FBI can now enter your home, search around, and doesn't ever have to tell you it was there. You could be perfectly innocent, yet federal agents can go through your most personal effects. When considering new laws, a test of the impact on liberty should be required. On that test, the Patriot Act fails. At a massive 342 pages, it potentially violates at least six of the ten original amendments known as the Bill of Rights — the First, Fourth, Fifth, Sixth, Seventh and Eighth Amendments — and possibly the Thirteenth and Fourteenth as well.

Please read the paragraphs carefully and, after each one, rate **the extent to which it decreases or increases your support for the Patriot Act**.

With the passage of the Patriot Act in 2001, the FBI can now enter your home, search around, and doesn't ever have to tell you it was there. You could be perfectly innocent, yet federal agents can go through your most personal effects. When considering new laws, a test of the impact on liberty should be required. On that test, the Patriot Act fails. At a massive 342 pages, it potentially violates at least six of the ten original amendments known as the Bill of Rights — the First, Fourth, Fifth, Sixth, Seventh and Eighth Amendments — and possibly the Thirteenth and Fourteenth as well.

Example[19]

Fears of Future Terror Attacks Warranted

By Andrew Tardaca
Published: January 17, 2009

U.S. citizens are bracing for another 9/11 type terrorist attack, according to a variety of reports. A recent Gallup poll finds that 87% of the American public is highly concerned about the possibility of a terrorist attack at home. According to new information from several international sources, these fears are well supported.

A raid on a London terrorist hideout on November 9, 2008 resulted in the capture of computer files that identified numerous U.S. financial districts, cultural centers, and transportation systems on a list of future Al Qaeda targets. According to a recent overseas intelligence report, "al Qaeda already has several cells operating in the U.S. that may be on the verge of mounting a large-scale terrorist attack."

On September 11, 2001, Al Qaeda's attacks killed nearly 3,000 men, women, and children, and injured over 6,000 more. Since September 11th, Al Qaeda and groups affiliated with Al Qaeda have waged attacks in countries such as Egypt, Indonesia, Kenya, Morocco, Saudi Arabia, Spain, Turkey, the United Kingdom, and most recently India. U.S. security officials are warning that current terrorist plots include plans for attacks on U.S. soil at least twice the magnitude of 9/11. An anonymous source reported that recent intelligence documents contain "sobering information" concerning the magnitude of future terrorist attacks.

Warnings issued by extremist groups such as Al Qaeda to "attack U.S. interests and allies on its soil" are even more alarming given the state of preparedness for future incidents. Experts have issued warnings about

[19] Merolla & Zechmeister. 2013. "Evaluating Political Leaders in Times of Terror and Economic Threat: The Conditioning Influence of Politician Partisanship." *Journal of Politics* 75(3): 599–712.

Example[19]

Economic Recession Projected to Deepen

By Andrew Tardaca
Published: January 17, 2009

U.S. citizens are bracing for a drastic deepening of the current economic recession. A recent Gallup poll finds that 87% of the American public is highly concerned about economic conditions in the country. The report further states "The economic mood is grimmer than it has been since 1992."

On September 16, failures of large financial institutions in the United States, such as Lehman Brothers and AIG, rapidly evolved into a global crisis resulting in bank failures across the U.S. and Europe. In the United States alone, 15 banks failed in 2008, while several others were rescued through government intervention or acquisitions by other banks. These events led to sharp reductions in the value of stocks and commodities worldwide. Over the past year, the Dow Jones Industrial Average lost 33.8%, the third worst loss in our nation's history. On October 11, 2008, the head of the International Monetary Fund (IMF) warned that the world financial system is teetering on the "brink of systemic meltdown".

The bank failures and subsequent market collapse were tied to sub-prime loans and credit default swaps. Increasing interest rates on loans hit the housing market particularly hard, as individuals were unable to keep up with mortgage payments. 2008 witnessed a record number of foreclosures, leading to the worst housing crisis, banking failure, and market collapse since the Great Depression.

Future projections are looking even grimmer. Experts predict that the housing market will not recover for at least a decade, especially now that banks are hesitant to make loans. The downturn in the economy has led to

[19] Merolla & Zechmeister. 2013. "Evaluating Political Leaders in Times of Terror and Economic Threat: The Conditioning Influence of Politician Partisanship." *Journal of Politics* 75(3): 599–712.

Some vignette considerations

Some vignette considerations

- Comparability across conditions
  - Length
  - Readability

Some vignette considerations

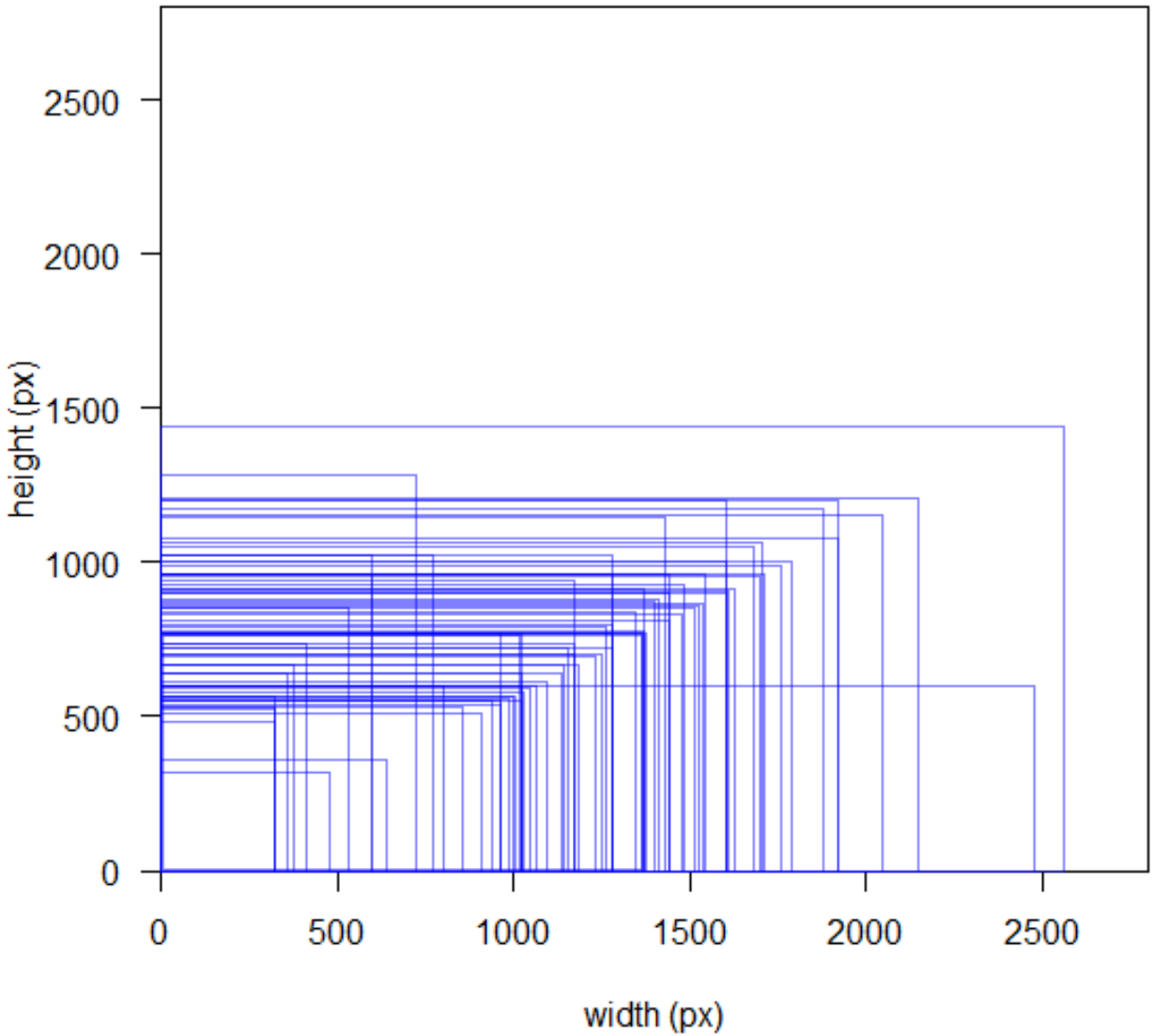
- Comparability across conditions
  - Length
  - Readability
- Language proficiency

Some vignette considerations

- Comparability across conditions
  - Length
  - Readability
- Language proficiency
- Length
  - Timers
  - Forced exposure
  - Mouse trackers

Some vignette considerations

- Comparability across conditions
  - Length
  - Readability
- Language proficiency
- Length
  - Timers
  - Forced exposure
  - Mouse trackers
- Devices
  - Browser-specificity
  - Device sizes (e.g., mobile)


2500
2000
1500
1000
500
0
height (px)
0
500
1000
1500
2000
2500
width (px)

Aside: Unique features of online studies

Aside: Unique features of online studies

- Capacity for audio-visual treatments and measurements

Aside: Unique features of online studies

- Capacity for audio-visual treatments and measurements
- Paradata collection
  - Implicit outcomes like response times, answer switching, mouse click behavior, browser focus, eye tracking, etc.

Aside: Unique features of online studies

- Capacity for audio-visual treatments and measurements
- Paradata collection
  - Implicit outcomes like response times, answer switching, mouse click behavior, browser focus, eye tracking, etc.
- Complex randomization

Aside: Unique features of online studies

- Capacity for audio-visual treatments and measurements
- Paradata collection
  - Implicit outcomes like response times, answer switching, mouse click behavior, browser focus, eye tracking, etc.
- Complex randomization
- Panel data
- Synchronous, multi-person designs

Non-textual Manipulations

- Images can work well
- Standalone or embedded in a text or question

[20] "Cueing Patriotism, Prejudice, and Partisanship in the Age of Obama: Experimental Tests of U.S. Flag Imagery Effects in Presidential Elections." *Political Psychology*: in press.

Non-textual Manipulations

- Images can work well
- Standalone or embedded in a text or question
- Examples
  - Kalmoe & Gross[20] measure impact of patriotic cues on candidate support by showing images of candidates with and without flags

[20] "Cueing Patriotism, Prejudice, and Partisanship in the Age of Obama: Experimental Tests of U.S. Flag Imagery Effects in Presidential Elections." *Political Psychology*: in press.

Non-textual Manipulations

- Images can work well
- Standalone or embedded in a text or question
- Examples
  - Kalmoe & Gross[20] measure impact of patriotic cues on candidate support by showing images of candidates with and without flags
  - Subliminal primes possible, depending on software

[20] "Cueing Patriotism, Prejudice, and Partisanship in the Age of Obama: Experimental Tests of U.S. Flag Imagery Effects in Presidential Elections." *Political Psychology*: in press.

Non-textual Manipulations

- Images can work well
- Standalone or embedded in a text or question
- Examples
  - Kalmoe & Gross[20] measure impact of patriotic cues on candidate support by showing images of candidates with and without flags
  - Subliminal primes possible, depending on software
  - Lots of recent examples of facial manipulation

[20] "Cueing Patriotism, Prejudice, and Partisanship in the Age of Obama: Experimental Tests of U.S. Flag Imagery Effects in Presidential Elections." *Political Psychology*: in press.

Example[21]

Light Complexion

Original

Dark Complexion

[21] Iyengar et al. 2010. "Do Explicit Racial Cues Influence Candidate Preference? The Case of Skin Complexion in the 2008 Campaign." Working paper.

Example[22]



[22] Laustsen & Petersen. 2016. "Winning Faces vary by Ideology." *Political Communication* 33(2): 188–211.

Example[23]

[23] Bailenson et al. 2006. "Transformed Facial Similarity as a Political Cue: A Preliminary Investigation." *Political Psychology* 27(3): 373–385.

Audio & Video manipulations

- Problematic for same reasons as long texts

[24] Vavreck. 2007 "The Exaggerated Effects of Advertising on Turnout: The Dangers of Self-Reports." *Quarterly Journal of Political Science* 2: 325–343.

[25] Mutz. 2007. "Effects of 'In-Your-Face' Television Discourse on Perceptions of a Legitimate Opposition." *American Political Science Review* 101(4): 621–635.

Audio & Video manipulations

- Problematic for same reasons as long texts
- Best practices
  - Keep it short
  - Have the video play automatically
  - Disallow survey progression
  - Control and validate

[24] Vavreck. 2007 "The Exaggerated Effects of Advertising on Turnout: The Dangers of Self-Reports." *Quarterly Journal of Political Science* 2: 325–343.

[25] Mutz. 2007. "Effects of 'In-Your-Face' Television Discourse on Perceptions of a Legitimate Opposition." *American Political Science Review* 101(4): 621–635.

Audio & Video manipulations

- Problematic for same reasons as long texts
- Best practices
  - Keep it short
  - Have the video play automatically
  - Disallow survey progression
  - Control and validate
- Examples
  - Television Advertisements[24]
  - News Programs[25]

---

[24] Vavreck. 2007 "The Exaggerated Effects of Advertising on Turnout: The Dangers of Self-Reports." *Quarterly Journal of Political Science* 2: 325–343.

[25] Mutz. 2007. "Effects of 'In-Your-Face' Television Discourse on Perceptions of a Legitimate Opposition." *American Political Science Review* 101(4): 621–635.

"Task" Designs

- Task designs ask respondents to perform a task
- Often developed for laboratory settings

"Task" Designs

- Task designs ask respondents to perform a task
- Often developed for laboratory settings
- Most common example: writing something

"Task" Designs

- Task designs ask respondents to perform a task
- Often developed for laboratory settings
- Most common example: writing something
- Can be problematic:
  - Time-intensive
  - Invites drop-off
  - Compliance problems

Univalent

These days, Democrats and Republicans differ from one another considerably. The two groups seem to be growing further and further apart, not only in terms of their opinions but also their lifestyles. Earlier in the survey, you said you tend to identify as a *Democrat/ Republican*. Please take a few minutes to think about what you like about *Democrats/ Republicans* compared to the *Republicans/ Democrats*. Think of 2 to 3 things you especially like best about **your party**. Then think of 2 to 3 things you especially dislike about **the other party**. Now please write those thoughts in the space below.

Ambivalent

These days, Democrats and Republicans differ from one another considerably. The two groups seem to be growing further and further apart, not only in terms of their opinions but also their lifestyles. Earlier in the survey, you said you tend to identify as a *Democrat/ Republican*. Please take a few minutes to think about what you like about *Democrats/ Republicans* compared to the *Republicans/ Democrats*. Think of 2 to 3 things you especially like best about **the other party**. Then think of 2 to 3 things you especially dislike about **your party**. Now please write those thoughts in the space below.

Questions?

Sensitive Item Designs

- Experiments can also be used to measure something
- Goal here is not necessarily causal inference, though the causal insight of the experiment provides *descriptively* useful information
- Paradigms
  - List experiments
  - Endorsement experiments

List Experiments [26]

Now I'm going to read you three things that sometimes make people angry or upset. After I read all three, just tell me *how many* of them upset you. I don't want to know which ones. just *how many*.

1. the federal government increasing the tax on gasoline
2. professional athletes getting million-dollar salaries
3. large corporations polluting the environment

[26] Kuklinski et al. 1997. "Racial Prejudice and Attitudes Toward Affirmative Action." *American Journal of Political Science* 41(2): 402–419.

List Experiments [26]

Now I'm going to read you three things that sometimes make people angry or upset. After I read all three, just tell me *how many* of them upset you. I don't want to know which ones. just *how many*.

1. the federal government increasing the tax on gasoline
2. professional athletes getting million-dollar salaries
3. large corporations polluting the environment
4. **a black family moving in next door**

[26] Kuklinski et al. 1997. "Racial Prejudice and Attitudes Toward Affirmative Action." *American Journal of Political Science* 41(2): 402–419.

Endorsement experiments[27]

A recent proposal calls for the sweeping reform of the Afghan prison system, including the construction of new prisons in every district to help alleviate overcrowding in existing facilities. Though expensive, new programs for inmates would also be offered, and new judges and prosecutors would be trained. How do you feel about this proposal?

[27] Lyall, Blair, & Imai. 2013. "Explaining Support for Combatants during Wartime: A Survey Experiment in Afghanistan." *American Political Science Review* 107(4): 679–705.

Endorsement experiments[27]

A recent proposal **by the Taliban** calls for the sweeping reform of the Afghan prison system, including the construction of new prisons in every district to help alleviate overcrowding in existing facilities. Though expensive, new programs for inmates would also be offered, and new judges and prosecutors would be trained. How do you feel about this proposal?

[27] Lyall, Blair, & Imai. 2013. "Explaining Support for Combatants during Wartime: A Survey Experiment in Afghanistan." *American Political Science Review* 107(4): 679–705.

Questions?

Let's work in R!

(Analysis of Example Experiments)

SUTO Framework

- Cronbach (1986) talks about generalizability in terms of UTO
- Shadish, Cook, and Campbell (2001) speak similarly of:
  - **S**ettings
  - **U**nits
  - **T**reatments
  - **O**utcomes
- External validity depends on all of these

Population

- Setting
- Units
- Treatments
- Outcomes

Your Study

- Setting
- Units
- Treatments
- Outcomes

Population

- Setting
- Units
- Treatments
- Outcomes

Your Study

- Setting
- Units
- Treatments
- Outcomes

In your study, how do these correspond?

Population

- Setting
- Units
- Treatments
- Outcomes

Your Study

- Setting
- Units
- Treatments
- Outcomes

In your study, how do these correspond?
how do these differ?

Population

- Setting
- Units
- Treatments
- Outcomes

Your Study

- Setting
- Units
- Treatments
- Outcomes

In your study, how do these correspond?
how do these differ?
do these differences matter?

Common Differences

- Most common thing to focus on is demographic representativeness
  - Sears (1986): "students aren't real people"
  - Western, educated, industrialized, rich, democratic (WEIRD) psychology participants

Common Differences

- Most common thing to focus on is demographic representativeness
  - Sears (1986): "students aren't real people"
  - Western, educated, industrialized, rich, democratic (WEIRD) psychology participants
- But do those characteristics actually matter?

Common Differences

- Most common thing to focus on is demographic representativeness
  - Sears (1986): "students aren't real people"
  - Western, educated, industrialized, rich, democratic (WEIRD) psychology participants
- But do those characteristics actually matter?
- Shadish, Cook, and Campbell tell us to think about:
  - Surface similarities
  - Ruling out irrelevancies
  - Making discriminations
  - Interpolation/extrapolation

Settings

- We should expect heterogeneity related to settings!
- How do we use/explore this?

Settings

- We should expect heterogeneity related to settings!
- How do we use/explore this?
  - Comparative research designs where experiments provide measures for each case

Settings

- We should expect heterogeneity related to settings!
- How do we use/explore this?
  - Comparative research designs where experiments provide measures for each case
  - Over-time replications of the same design

Settings

- We should expect heterogeneity related to settings!
- How do we use/explore this?
  - Comparative research designs where experiments provide measures for each case
  - Over-time replications of the same design
  - Replication of a design across contexts with unknown sources of variability?

Settings

- We should expect heterogeneity related to settings!
- How do we use/explore this?
  - Comparative research designs where experiments provide measures for each case
  - Over-time replications of the same design
  - Replication of a design across contexts with unknown sources of variability?
- Can we control for context?

Pretreatment Dynamics

"If the experiment explores a communication that regularly occurs in 'reality,' then reactions in the experiment might be contaminated by those 'regular' occurrences prior to the experiment." [28]

[28] p.875 from Druckman & Leeper. 2012. "Learning More from Political Communication Experiments: Pretreatment and Its Effects." *American Journal of Political Science* 56(4): 875–896.

Pretreatment Dynamics

- Pretreatment is a feature of an experimental setting, treatment, and sample, wherein the effect of the treatment has already occurred[29]

[29] Or, units having already been treated are otherwise affected differently.

Pretreatment Dynamics

- Pretreatment is a feature of an experimental setting, treatment, and sample, wherein the effect of the treatment has already occurred[29]
- Consequences:
  - Biased effect estimates

[29] Or, units having already been treated are otherwise affected differently.

Pretreatment Dynamics

- Pretreatment is a feature of an experimental setting, treatment, and sample, wherein the effect of the treatment has already occurred[29]
- Consequences:
  - Biased effect estimates
- Mitigation:
  - Measure pretreatment
  - Avoid "pretreated" treatments or contexts
  - Study units not already treated
  - Theorize repeated effects

[29] Or, units having already been treated are otherwise affected differently.

Questions?

Units

Most commonly studied source of heterogeneity is covariate-related (i.e., characteristics of units).

If we think there might be covariate-related effect heterogeneity, what can we do?

- Best solution: manipulate the moderator
- Next best: block on the moderator
- Least best: post-hoc exploratory approaches

Block Randomization I

Stratification:Sampling::Blocking:Experiments

Block Randomization I

Stratification:Sampling::Blocking:Experiments

- Basic idea: randomization occurs within strata defined before treatment assignment

Block Randomization I

Stratification:Sampling::Blocking:Experiments

- Basic idea: randomization occurs within strata defined before treatment assignment
- CATE is estimate for each stratum; aggregated to SATE

Block Randomization I

Stratification:Sampling::Blocking:Experiments

- Basic idea: randomization occurs within strata defined before treatment assignment
- CATE is estimate for each stratum; aggregated to SATE
- Why?
  - Eliminate chance imbalances
  - Optimized for estimating CATEs
  - More precise SATE estimate

Exp.	Control				Treatment			
1	M	M	M	M	F	F	F	F
2	M	M	M	F	M	F	F	F
3	M	M	F	F	M	M	F	F
4	M	F	F	F	M	M	M	F
5	F	F	F	F	M	M	M	M

```
# population of men and women
pop <- rep(c("Male", "Female"), each = 4)

# randomly assign into treatment and control
split(sample(pop, 8, FALSE), c(rep(0,4), rep(1,4)))
```

Obs.	X_{1i}	X_{2i}	D_i
1	Male	Old	0
2	Male	Old	1
3	Male	Young	1
4	Male	Young	0
5	Female	Old	1
6	Female	Old	0
7	Female	Young	0
8	Female	Young	1

Block Randomization II

- Blocking ensures ignorability of all covariates used to construct the blocks
- Incorporates covariates explicitly into the *design*

Block Randomization II

- Blocking ensures ignorability of all covariates used to construct the blocks
- Incorporates covariates explicitly into the *design*
- When is blocking *statistically* useful?

Block Randomization II

- Blocking ensures ignorability of all covariates used to construct the blocks
- Incorporates covariates explicitly into the *design*
- When is blocking *statistically* useful?
  - If those covariates affect values of potential outcomes, blocking reduces the variance of the SATE

Block Randomization II

- Blocking ensures ignorability of all covariates used to construct the blocks
- Incorporates covariates explicitly into the *design*
- When is blocking *statistically* useful?
  - If those covariates affect values of potential outcomes, blocking reduces the variance of the SATE
  - Most valuable in small samples

Block Randomization II

- Blocking ensures ignorability of all covariates used to construct the blocks
- Incorporates covariates explicitly into the *design*
- When is blocking *statistically* useful?
  - If those covariates affect values of potential outcomes, blocking reduces the variance of the SATE
  - Most valuable in small samples
  - Not valuable if all blocks have similar potential outcomes

Statistical Properties I

Complete randomization:

$$SATE = \frac{1}{n_1} \sum Y_{1i} - \frac{1}{n_0} \sum Y_{0i}$$

Block randomization:

$$SATE_{blocked} = \sum_{1}^{J} \left(\frac{n_j}{n} \right) (\widehat{CATE}_j)$$

Obs.	X_{1i}	X_{2i}	D_i	Y_i	CATE
1	Male	Old	0	5	
2	Male	Old	1	10	
3	Male	Young	1	4	
4	Male	Young	0	1	
5	Female	Old	1	6	
6	Female	Old	0	2	
7	Female	Young	0	6	
8	Female	Young	1	9	

Obs.	X_{1i}	X_{2i}	D_i	Y_i	CATE
1	Male	Old	0	5	5
2	Male	Old	1	10	
3	Male	Young	1	4	
4	Male	Young	0	1	
5	Female	Old	1	6	
6	Female	Old	0	2	
7	Female	Young	0	6	
8	Female	Young	1	9	

Obs.	X_{1i}	X_{2i}	D_i	Y_i	CATE
1	Male	Old	0	5	5
2	Male	Old	1	10	
3	Male	Young	1	4	3
4	Male	Young	0	1	
5	Female	Old	1	6	
6	Female	Old	0	2	
7	Female	Young	0	6	
8	Female	Young	1	9	

Obs.	X_{1i}	X_{2i}	D_i	Y_i	CATE
1	Male	Old	0	5	5
2	Male	Old	1	10	
3	Male	Young	1	4	3
4	Male	Young	0	1	
5	Female	Old	1	6	4
6	Female	Old	0	2	
7	Female	Young	0	6	
8	Female	Young	1	9	

Obs.	X_{1i}	X_{2i}	D_i	Y_i	CATE
1	Male	Old	0	5	5
2	Male	Old	1	10	
3	Male	Young	1	4	3
4	Male	Young	0	1	
5	Female	Old	1	6	4
6	Female	Old	0	2	
7	Female	Young	0	6	3
8	Female	Young	1	9	

SATE Estimation

$$\begin{aligned} SATE &= \left(\frac{2}{8} * 5\right) + \left(\frac{2}{8} * 3\right) + \left(\frac{2}{8} * 4\right) + \left(\frac{2}{8} * 3\right) \\ &= 3.75 \end{aligned}$$

SATE Estimation

$$\begin{aligned} SATE &= \left(\frac{2}{8} * 5\right) + \left(\frac{2}{8} * 3\right) + \left(\frac{2}{8} * 4\right) + \left(\frac{2}{8} * 3\right) \\ &= 3.75 \end{aligned}$$

The blocked and unblocked estimates are the same here because $Pr(Treatment)$ is constant across blocks and blocks are all the same size.

SATE Estimation

- We can use weighted regression to estimate this in an OLS framework
- Weights are the inverse prob. of being treated w/in block
  - Pr(Treated) by block: $p_{ij} = Pr(D_i = 1|J = j)$
  - Weight (Treated): $w_{ij} = \frac{1}{p_{ij}}$
  - Weight (Control): $w_{ij} = \frac{1}{1 - p_{ij}}$

Statistical Properties II

Complete randomization:

$$\widehat{SE}_{SATE} = \sqrt{\frac{\widehat{Var}(Y_0)}{n_0} + \frac{\widehat{Var}(Y_1)}{n_1}}$$

Block randomization:

$$\widehat{SE}_{SATE_{blocked}} = \sqrt{\sum_{1}^{J} \left(\frac{n_j}{n}\right)^2 \widehat{Var}(SATE_j)}$$

Statistical Properties II

Complete randomization:

$$\widehat{SE}_{SATE} = \sqrt{\frac{\widehat{Var}(Y_0)}{n_0} + \frac{\widehat{Var}(Y_1)}{n_1}}$$

Block randomization:

$$\widehat{SE}_{SATE_{blocked}} = \sqrt{\sum_{1}^{J} \left(\frac{n_j}{n}\right)^2 \widehat{Var}(SATE_j)}$$

When is the blocked design more efficient?

Practicalities

- Blocked randomization only works in exactly the same situations where stratified sampling works
  - Need to observe covariates pre-treatment in order to block on them
  - Work best in a panel context
- In a single cross-sectional design that might be challenging
  - Some software can block "on the fly"

Questions?

Three Post-hoc Approaches

- Suggestive evidence
- Regression using treatment-by-covariate interactions
- Automated approaches

Three Post-hoc Approaches

- Suggestive evidence
- Regression using treatment-by-covariate interactions
- Automated approaches
- (Replication and meta-analysis)

Suggestive Evidence

We can never know $Var(TE_i)$!

Suggestive Evidence

We can never know $Var(TE_i)$! But...

- Quantile-quantile plots
- Equality of variance tests

Suggestive Evidence

We can never know $Var(TE_i)$! But...

- Quantile-quantile plots
  - Compare the distribution of Y_0's to distribution of Y_1's
  - If homogeneity, a vertical shift in Y_1's
  - If heterogeneity, a slope $\neq 1$
- Equality of variance tests

Suggestive Evidence

We can never know $Var(TE_i)$! But...

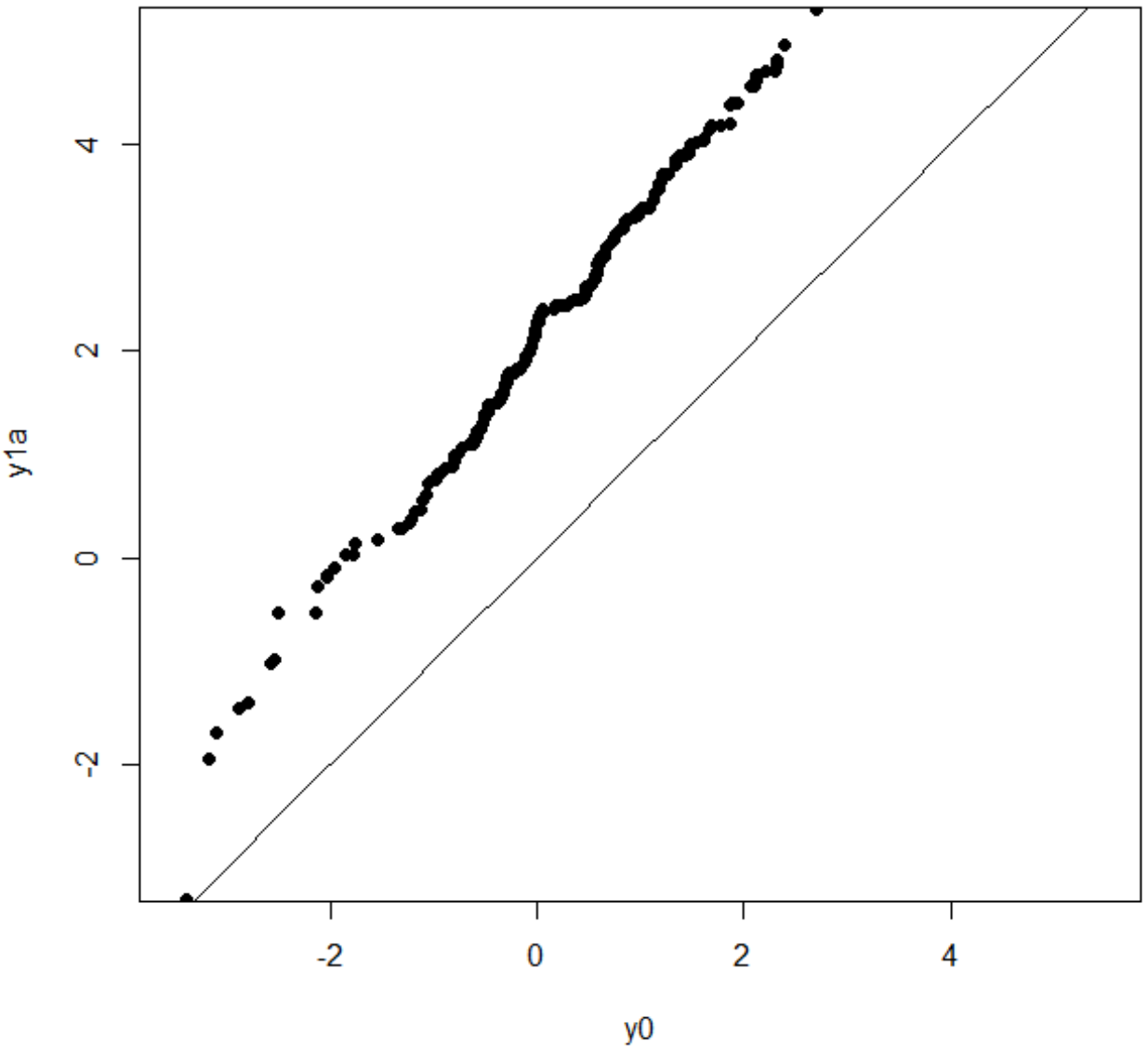
- Quantile-quantile plots
  - Compare the distribution of Y_0's to distribution of Y_1's
  - If homogeneity, a vertical shift in Y_1's
  - If heterogeneity, a slope $\neq 1$
- Equality of variance tests
  - If homogeneity, variance should be equal
  - If heterogeneity, variances should differ

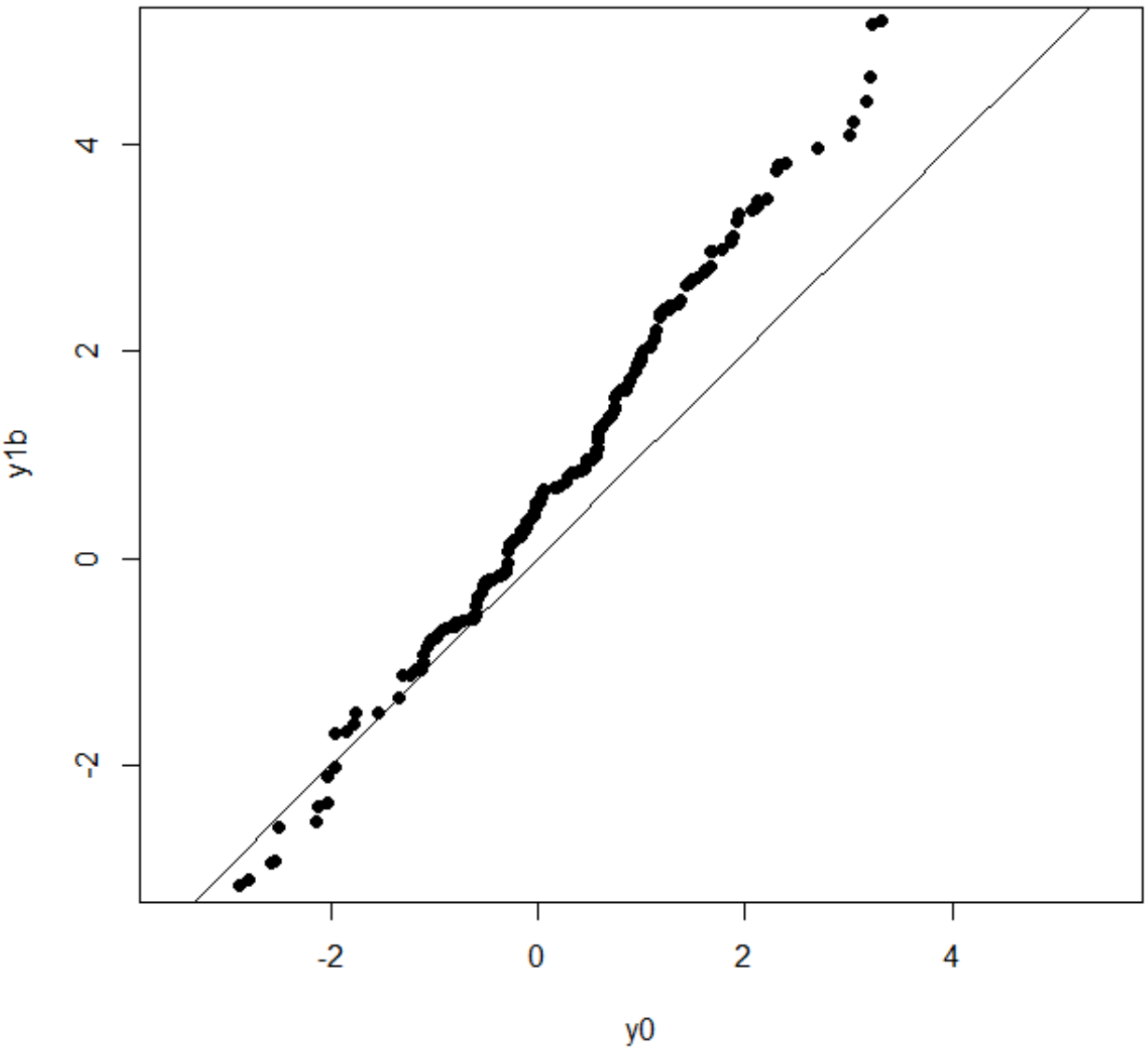
QQ Plots

```
# y_0 data
set.seed(1)
n <- 200
y0 <- rnorm(n) + rnorm(n, 0.2)

# y_1 data (homogeneous effects)
y1a <- y0 + 2 + rnorm(n, 0.2)
# y_1 data (heterogeneous effects)
y1b <- y0 + rep(0:1, each = n/2) + rnorm(n, 0.2)

qqplot(y0, y1a, pch=19, xlim=c(-3,5), ylim=c(-3,5), asp=1)
curve((x), add = TRUE)
qqplot(y0, y1b, pch=19, xlim=c(-3,5), ylim=c(-3,5), asp=1)
curve((x), add = TRUE)
```


y1a
4
2
0
-2
-2
0
2
4
y0


y1b
4
2
0
-2
-2
0
2
4
y0

Equality of Variance tests

```
> var.test(y0, y1a)

        F test to compare two variances

data:  y0 and y1a
F = 0.60121, num df = 199, denom df = 199,
  p-value = 0.0003635
alternative hypothesis:
  true ratio of variances is not equal to 1
95 percent confidence interval:
 0.4549900 0.7944289
sample estimates:
ratio of variances
         0.6012131
```

Equality of Variance tests

```
> var.test(y0, y1b)

        F test to compare two variances

data:  y0 and y1b
F = 0.53483, num df = 199, denom df = 199,
  p-value = 1.224e-05
alternative hypothesis:
  true ratio of variances is not equal to 1
95 percent confidence interval:
 0.4047531 0.7067133
sample estimates:
ratio of variances
         0.5348312
```

Questions?

Regression Estimation

Aside: Regression Adjustment in Experiments, Generally

- Recall the general advice that we do not need covariates in the regression to "control" for omitted variables (because there are none)
- Including covariates can reduce variance of our SATE by explaining more of the variation in Y

Scenario

Imagine two regression models. Which is correct?

1. Mean-difference estimate of SATE is "not significant"
2. Regression estimate of SATE, controlling for sex, age, and education, is "significant"

Scenario

Imagine two regression models. Which is correct?

1. Mean-difference estimate of SATE is "not significant"
2. Regression estimate of SATE, controlling for sex, age, and education, is "significant"

This is a small-sample dynamic, so make these decisions pre-analysis!

Treatment-Covariate Interactions

- The regression paradigm allows us to estimate CATEs using interaction terms
  - X is an indicator for treatment
  - M is an indicator for possible moderator

Treatment-Covariate Interactions

- The regression paradigm allows us to estimate CATEs using interaction terms
  - X is an indicator for treatment
  - M is an indicator for possible moderator
- SATE: $Y = \beta_0 + \beta_1 X + e$

Treatment-Covariate Interactions

- The regression paradigm allows us to estimate CATEs using interaction terms
  - X is an indicator for treatment
  - M is an indicator for possible moderator
- SATE: $Y = \beta_0 + \beta_1 X + e$
- CATEs:

$$Y = \beta_0 + \beta_1 X + \beta_2 M + \beta_3 X * M + e$$

Treatment-Covariate Interactions

- The regression paradigm allows us to estimate CATEs using interaction terms
  - X is an indicator for treatment
  - M is an indicator for possible moderator
- SATE: $Y = \beta_0 + \beta_1 X + e$
- CATEs:

$$Y = \beta_0 + \beta_1 X + \beta_2 M + \beta_3 X * M + e$$

  - Homogeneity: $\beta_3 = 0$
  - Heterogeneity: $\beta_3 \neq 0$

Let's work in R!

(Covariate-related effect heterogeneity)

BART

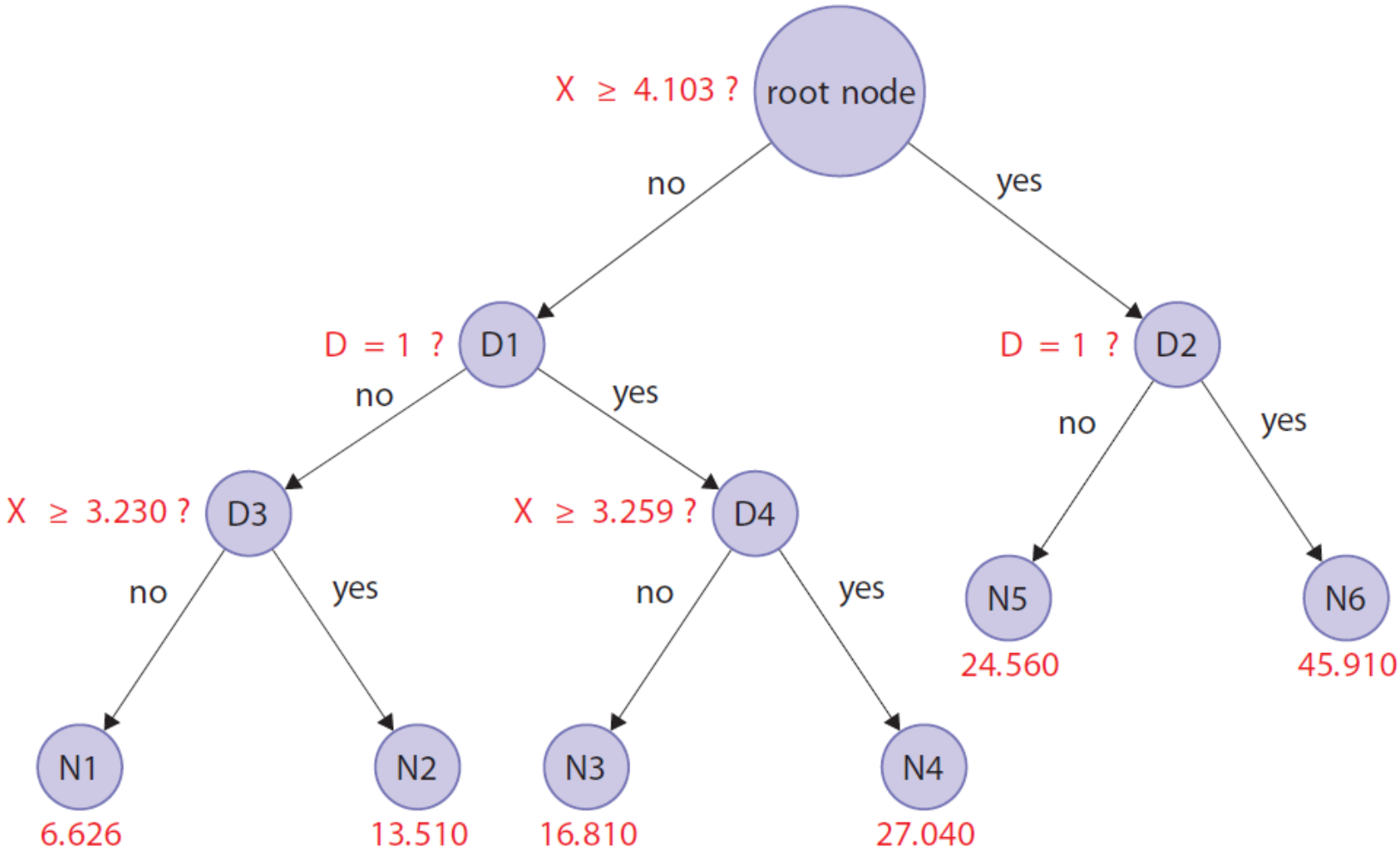
- Estimate CATEs in a fully automated fashion

BART

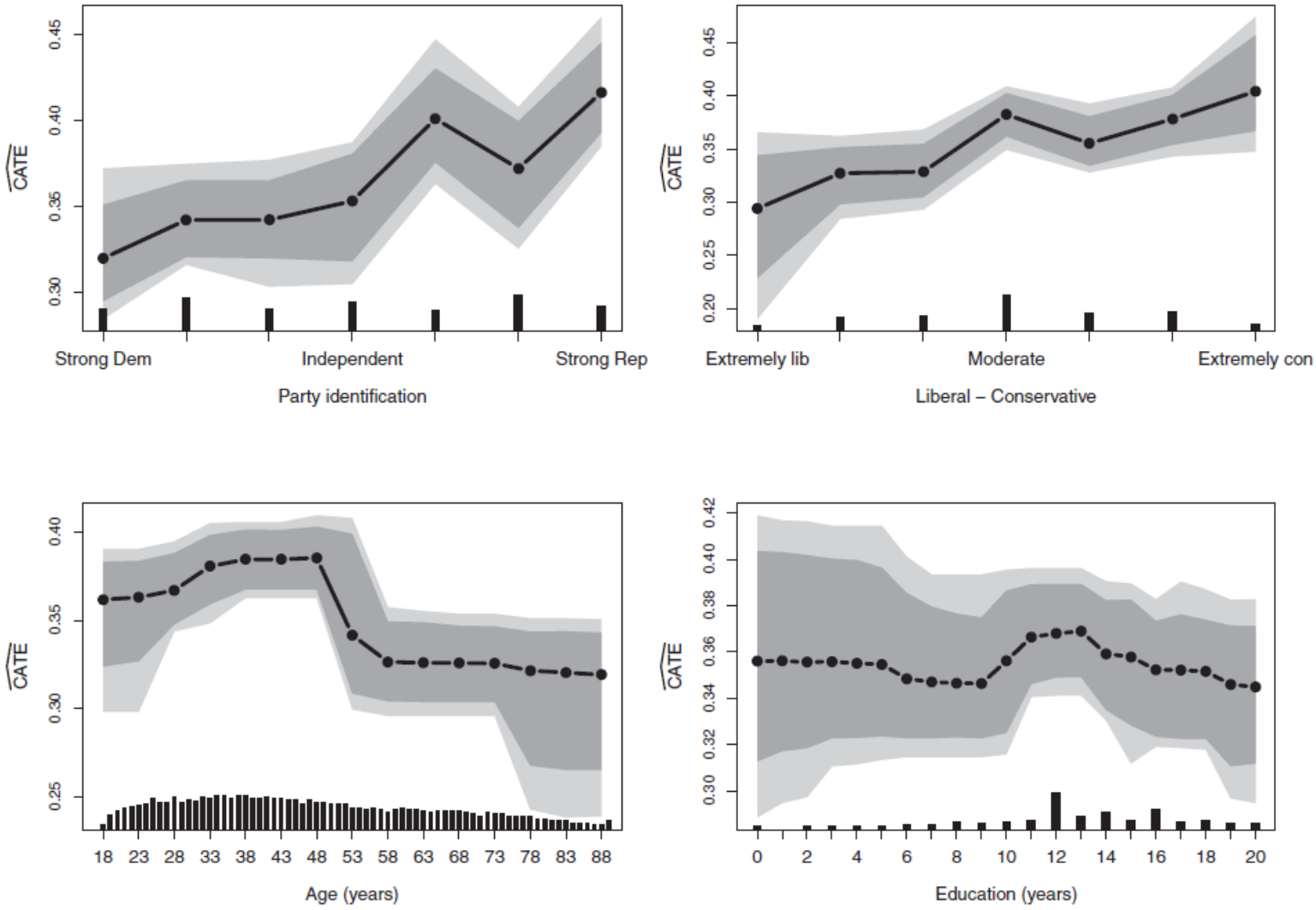
- Estimate CATEs in a fully automated fashion
- "Bayesian Additive Regression Trees"
  - Essentially an ensemble machine learning method

BART

- Estimate CATEs in a fully automated fashion
- "Bayesian Additive Regression Trees"
  - Essentially an ensemble machine learning method
- Iteratively split a sample into more and more homogeneous groups until some threshold is reached using binary (cutpoint) decisions
- Repeat this a bunch of times, aggregating across results



Green & Kern. 2012. "Modeling Heterogeneous Treatment Effects in Survey Experiments with Bayesian Additive Regression Trees." *Public Opinion Quarterly* 76(3): 491–511.


CATE
0.30
0.35
0.40
0.45
Strong Dem
Independent
Strong Rep
Party identification
CATE
0.20
0.25
0.30
0.35
0.40
0.45
Extremely lib
Moderate
Extremely con
Liberal – Conservative
CATE
0.25
0.30
0.35
0.40
18 23 28 33 38 43 48 53 58 63 68 73 78 83 88
Age (years)
CATE
0.30
0.32
0.34
0.36
0.38
0.40
0.42
0 2 4 6 8 10 12 14 16 18 20
Education (years)

Considerations

- BART is totally automated, conditional on the set of covariates used
- Only really works with dichotomous covariates
- Not widely used or tested
- Totally post-hoc and atheoretical

Considerations

Considerations

- Coefficients on moderators have no causal interpretation without further conditioning on observables

Considerations

- Coefficients on moderators have no causal interpretation without further conditioning on observables
- Nearly unlimited potential moderators
  - First-order interactions with every covariate in dataset
  - Second-, third-order, etc. interactions
- Thus, multiple comparisons problem!

Considerations

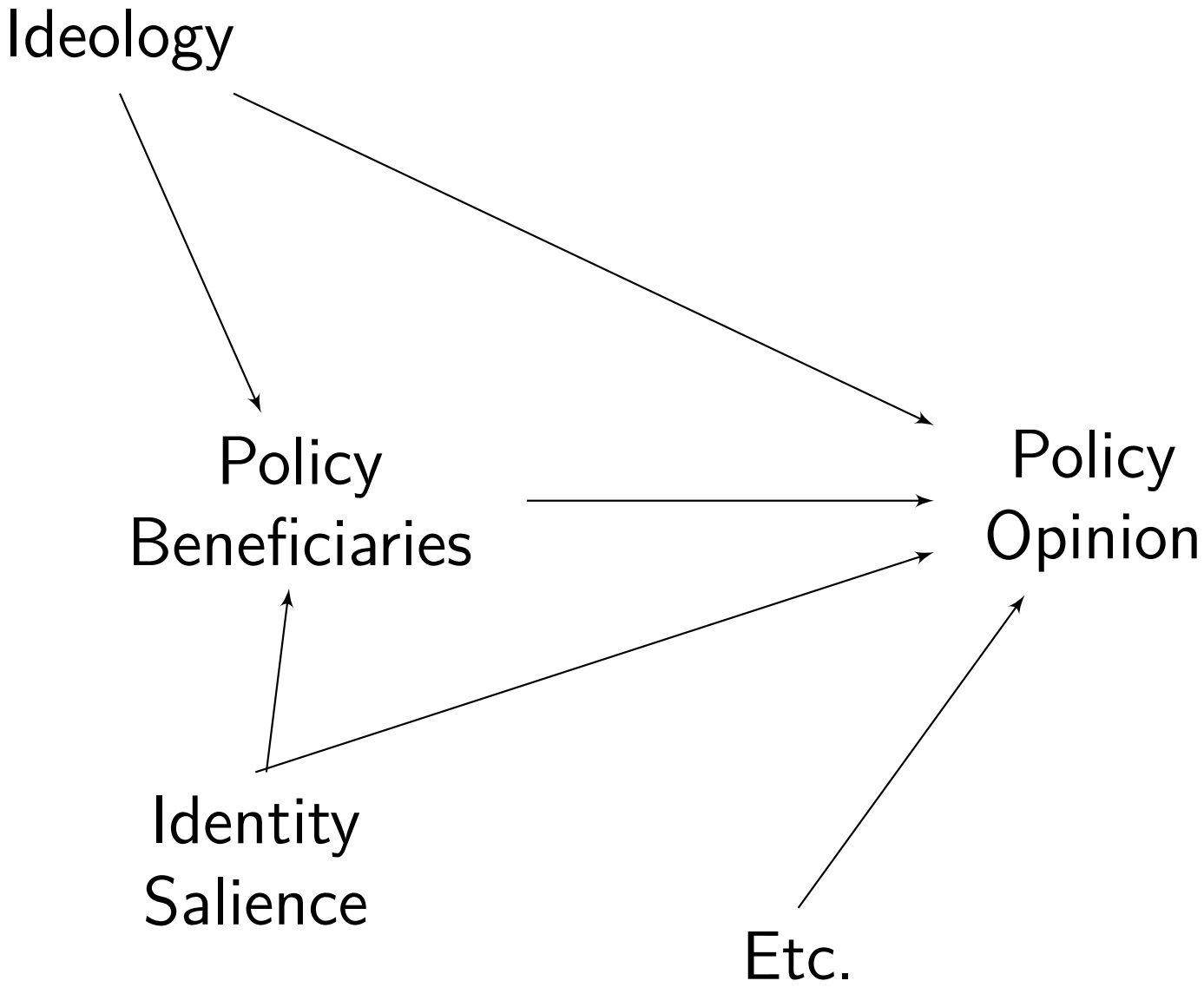
- Coefficients on moderators have no causal interpretation without further conditioning on observables
- Nearly unlimited potential moderators
  - First-order interactions with every covariate in dataset
  - Second-, third-order, etc. interactions
- Thus, multiple comparisons problem!
- Power (esp. if M is continuous)

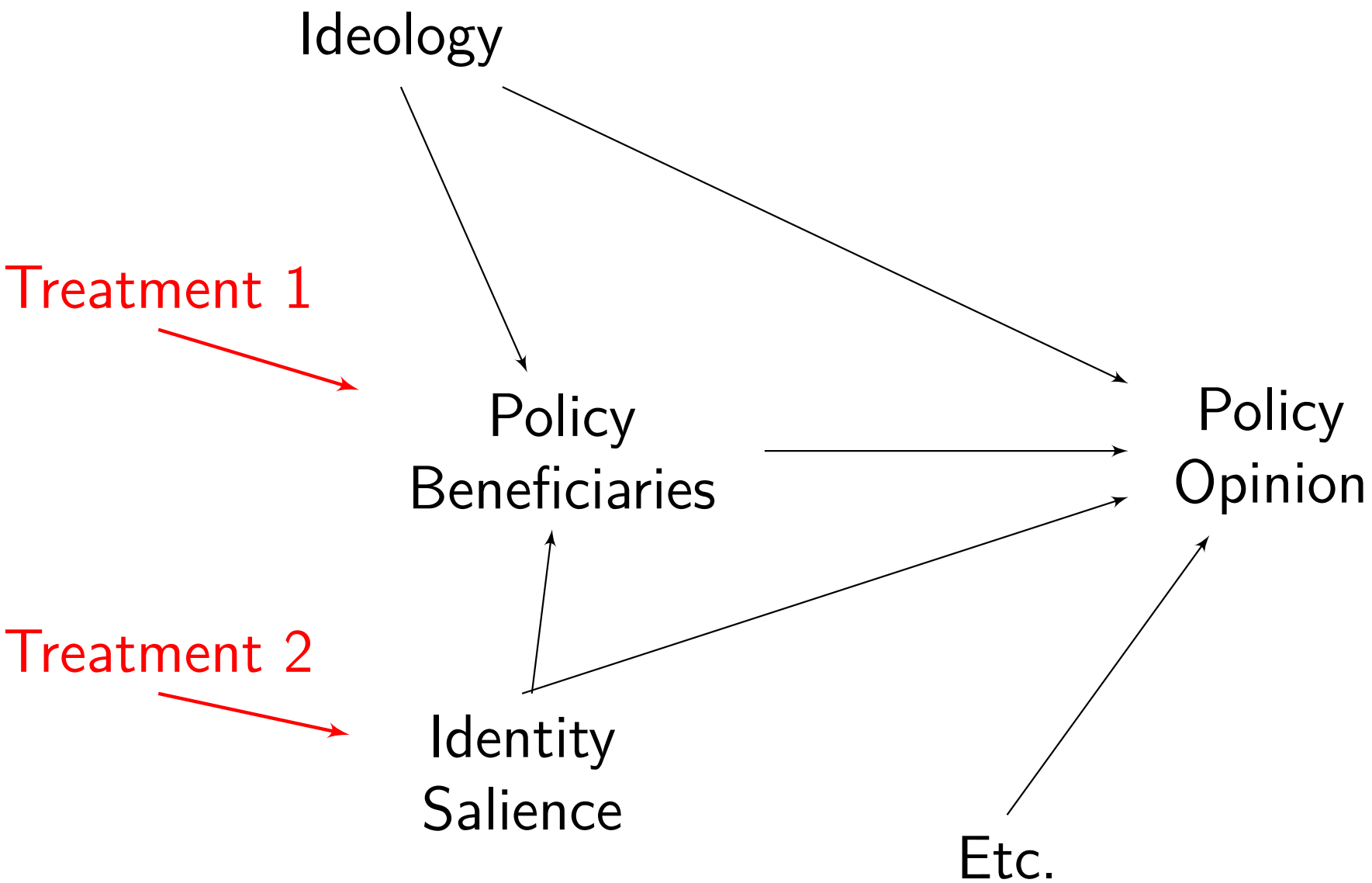
Simply: Manipulating the moderator variable is the best way to estimate a heterogeneous effect!

Why is this true?

Complex Designs

- An experiment can have any number of conditions
  - Up to the limits of sample size
  - More than 8–10 conditions is typically unwieldy
- Typically analyze complex designs using ANOVA or regression, but we are still ultimately interested in pairwise comparisons to estimates SATEs
  - Treatment–treatment, or treatment-control
  - Without control group, we don't know which treatment(s) affected the outcome


Ideology
Policy Beneficiaries
Policy Opinion
Identity Salience
Etc.


Ideology
Treatment 1
Policy
Beneficiaries
Policy
Opinion
Treatment 2
Identity
Salience
Etc.

Ex. Question-as-treatment[30]

- How close do you feel to your ethnic or racial group?
- Some people have said that taxes need to be raised to take care of pressing national needs. How willing would you be to have your taxes raised to improve education in public schools?

[30] Transue. 2007. "Identity Salience, Identity Acceptance, and Racial Policy Attitudes: American National Identity as a Uniting Force." *American Journal of Political Science* 51(1): 78–91.

Ex. Question-as-treatment[30]

- How close do you feel to other Americans?
- Some people have said that taxes need to be raised to take care of pressing national needs. How willing would you be to have your taxes raised to improve education in public schools?

[30] Transue. 2007. "Identity Salience, Identity Acceptance, and Racial Policy Attitudes: American National Identity as a Uniting Force." *American Journal of Political Science* 51(1): 78–91.

Ex. Question-as-treatment[30]

- How close do you feel to your ethnic or racial group?
- Some people have said that taxes need to be raised to take care of pressing national needs. How willing would you be to have your taxes raised to improve educational opportunities for minorities?

[30] Transue. 2007. "Identity Salience, Identity Acceptance, and Racial Policy Attitudes: American National Identity as a Uniting Force." *American Journal of Political Science* 51(1): 78–91.

Ex. Question-as-treatment[30]

- How close do you feel to other Americans?
- Some people have said that taxes need to be raised to take care of pressing national needs. How willing would you be to have your taxes raised to improve educational opportunities for minorities?

[30] Transue. 2007. "Identity Salience, Identity Acceptance, and Racial Policy Attitudes: American National Identity as a Uniting Force." *American Journal of Political Science* 51(1): 78–91.

2x2 Factorial Design

Condition	
Educ. for Minorities	Y_1
Schools	Y_0

2x2 Factorial Design

Condition	Americans	Own Race
Educ. for Minorities	$Y_{1,0}$	$Y_{1,1}$
Schools	$Y_{0,0}$	$Y_{0,1}$

Two ways to estimate this

Dummy variable regression:

$$Y = \beta_0 + \beta_1 X_{0,1} + \beta_2 X_{1,0} + \beta_3 X_{1,1} + \epsilon$$

Interaction effect:

$$Y = \beta_0 + \beta_1 X1_1 + \beta_2 X2_1 + \beta_3 X1_1 * X2_1 + \epsilon$$

Considerations

- Need to have hypotheses about heterogeneity a priori
- Factorial designs can quickly become unwieldy and expensive

Probably obvious, but. . .

Factors	Conditions per factor	Total Conditions	n
1	2	2	400
1	3	3	600
1	4	4	800
2	2	4	800
2	3	6	1200
2	4	8	1600
3	3	9	1800
3	4	12	2400
4	4	16	3200

Assumes power to detect a relatively small effect, but no consideration of multiple comparisons.

Considerations

- Need to have hypotheses about heterogeneity a priori
- Factorial designs can quickly become unwieldy and expensive

Considerations

- Need to have hypotheses about heterogeneity a priori
- Factorial designs can quickly become unwieldy and expensive
- Need to consider what CATEs are of theoretical interest
  - Treatment–control
  - Treatment–treatment

Questions?

One final issue with unit-related sources of heterogeneity is how we handle or analyze survey-experimental data where we think participants misbehaved.

One final issue with unit-related sources of heterogeneity is how we handle or analyze survey-experimental data where we think participants misbehaved.

This falls into a couple of broad categories:

- Noncompliance
- Inattention
- Survey Satisficing

How should we deal with respondents that appear to not be paying attention, not "taking" the treatment, or not responding to outcome measures?

1. Keep them
2. Throw them away

Best Practice: Protocol

- Excluding respondents based on survey behavior is one of the easiest ways to "p-hack" an experimental dataset
  - Inattention, satisficing, etc. will tend to reduce the size of the SATE
- So regardless of how you handle these respondents, these should be decisions that are made *pre-analysis*

When are you excluding participants?

Pre-Treatment

Post-Treatment

When are you excluding participants?

Pre-Treatment

- Satisficing behaviors

Post-Treatment

When are you excluding participants?

Pre-Treatment

- Satisficing behaviors
- Inattention

Post-Treatment

When are you excluding participants?

Pre-Treatment

- Satisficing behaviors
- Inattention
- Covariate-based selection

Post-Treatment

When are you excluding participants?

Pre-Treatment

- Satisficing behaviors
- Inattention
- Covariate-based selection
- Pretreated

Post-Treatment

When are you excluding participants?

Pre-Treatment

- Satisficing behaviors
- Inattention
- Covariate-based selection
- Pretreated

Post-Treatment

- Speeding on treatment

When are you excluding participants?

Pre-Treatment

- Satisficing behaviors
- Inattention
- Covariate-based selection
- Pretreated

Post-Treatment

- Speeding on treatment
- "Failing" a manipulation check

When are you excluding participants?

Pre-Treatment

- Satisficing behaviors
- Inattention
- Covariate-based selection
- Pretreated

Post-Treatment

- Speeding on treatment
- "Failing" a manipulation check
- Drop-off

Pre-Treatment Exclusion

- This is totally fine from a causal inference perspective

Pre-Treatment Exclusion

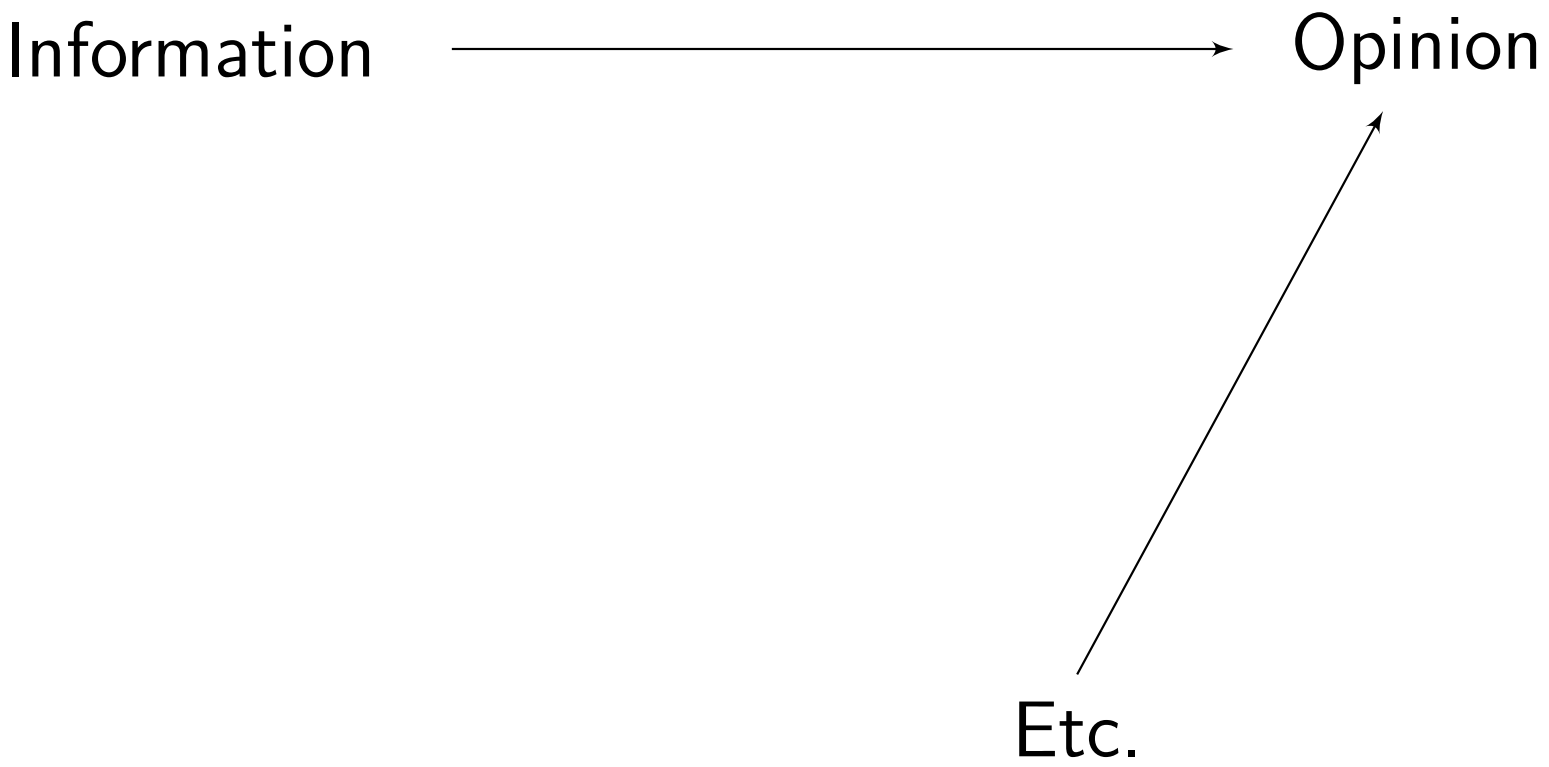
- This is totally fine from a causal inference perspective
- Advantages:
  - Focused on engaged respondents
  - Likely increase impact of treatment

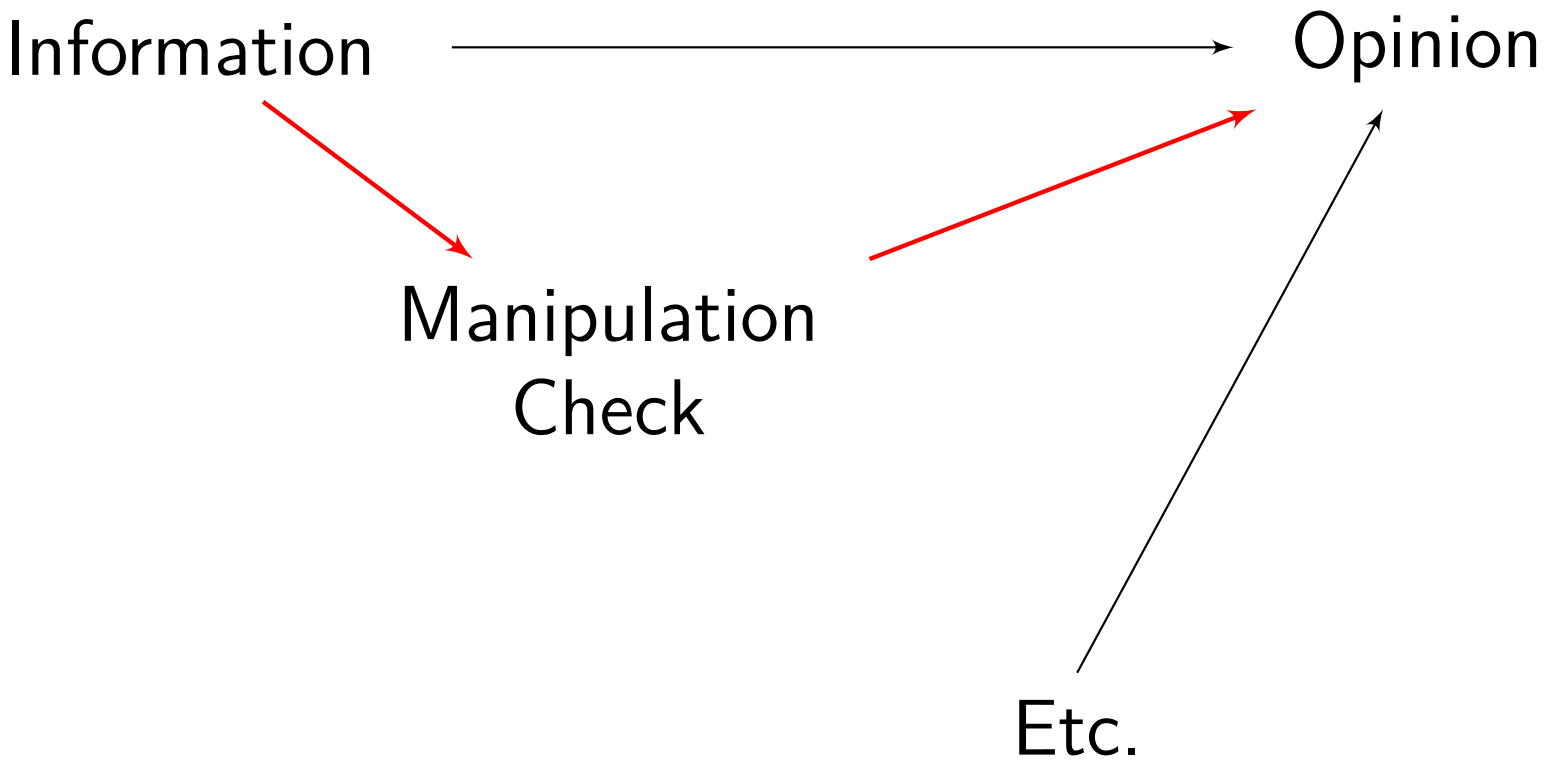
Pre-Treatment Exclusion

- This is totally fine from a causal inference perspective
- Advantages:
  - Focused on engaged respondents
  - Likely increase impact of treatment
- Disadvantages:
  - Changing definition of sample (and thus population)

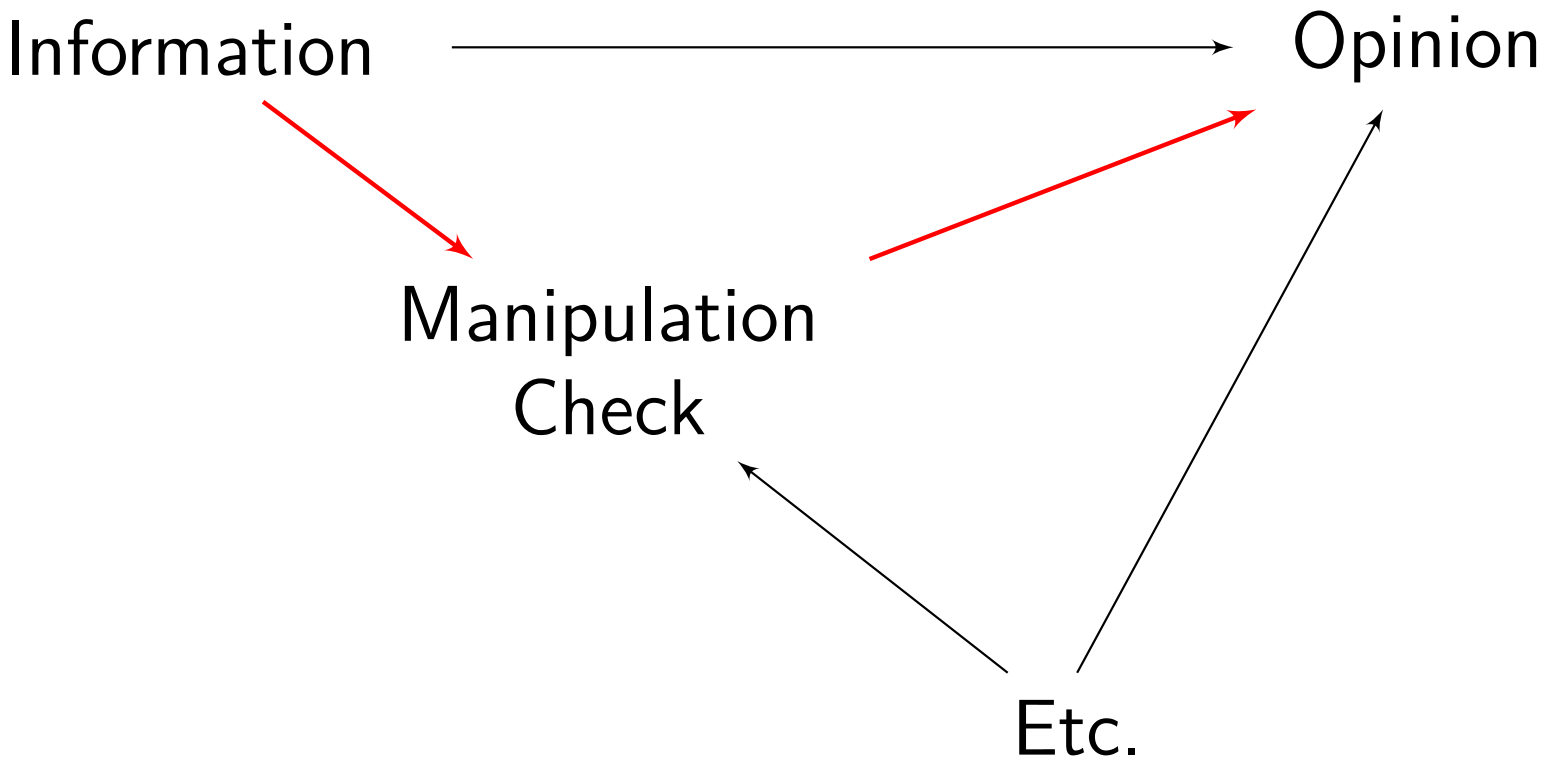
Post-Treatment Exclusion

This is much more problematic because it involves controlling for a *post-treatment* variable


Information
Opinion
Etc.



Risk that estimate of β_1 is diminished because effect is being carried through the manipulation check.



Introduction of "collider bias" wherein values of the manipulation check are affected by other factors.

Post-Treatment Exclusion

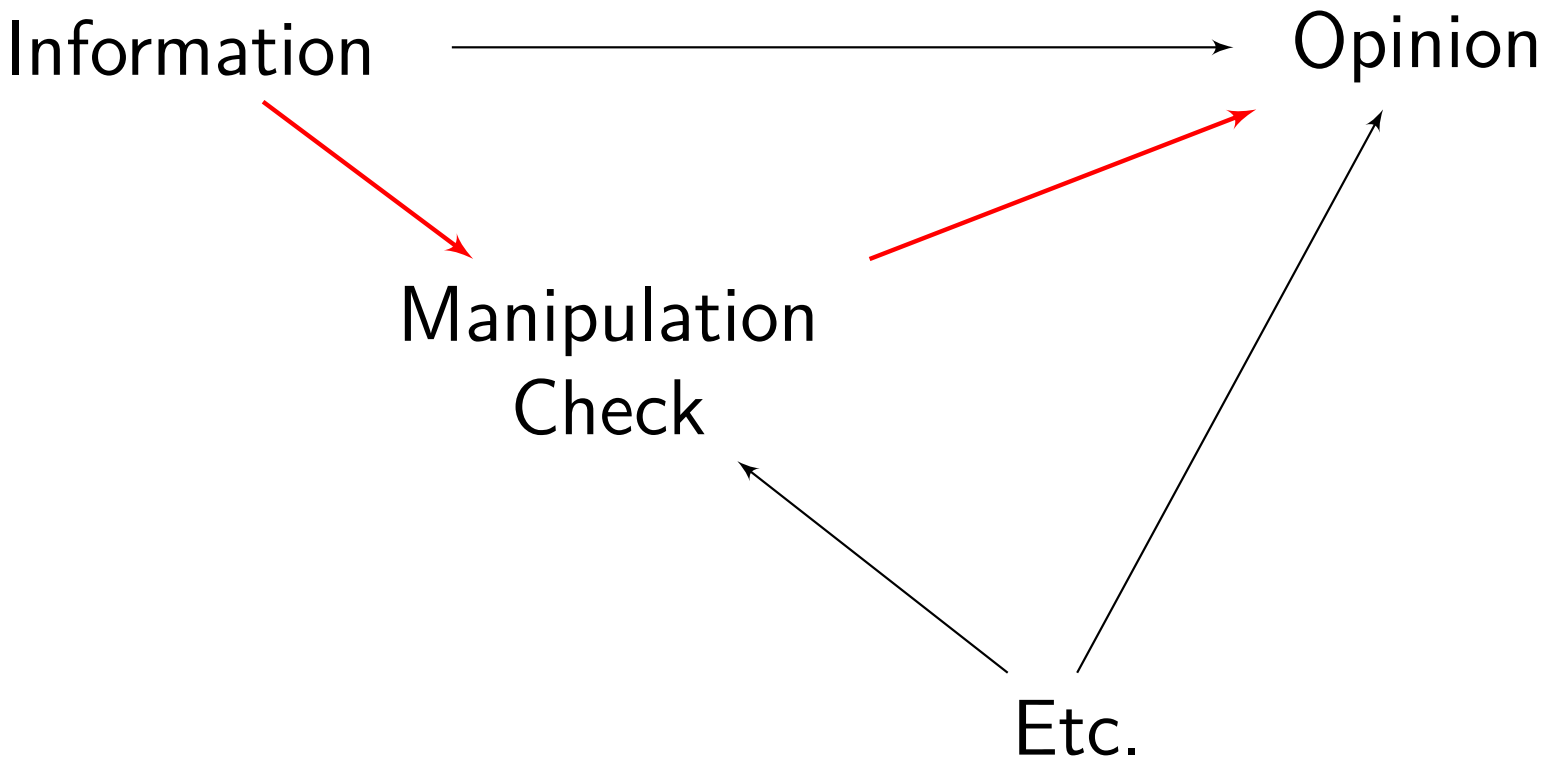
- Any post-treatment exclusion is problematic and should be avoided

Post-Treatment Exclusion

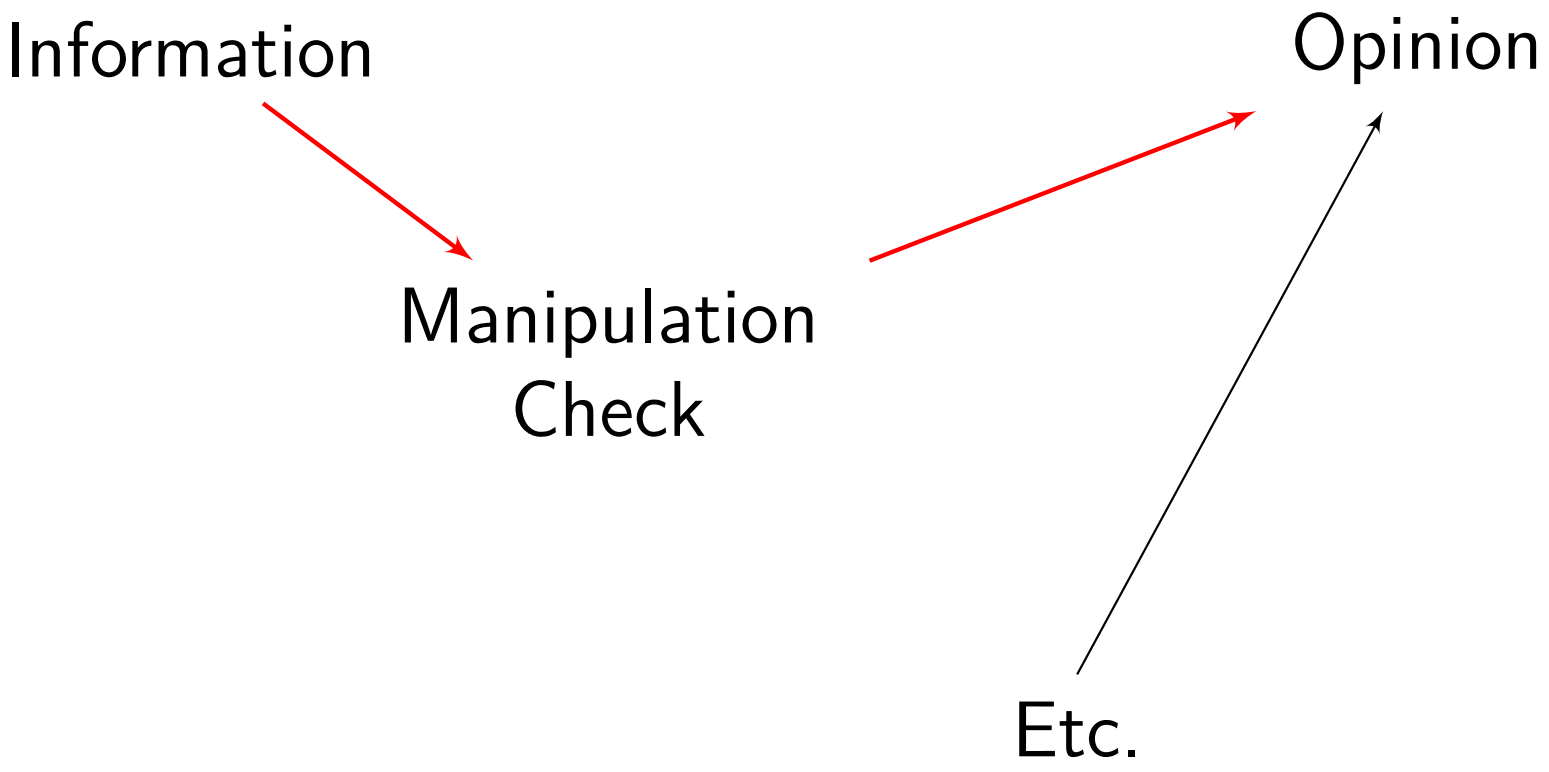
- Any post-treatment exclusion is problematic and should be avoided
- Can estimate a LATE
  - Interpretation: Effect of manipulation check among those whose value of the check can be changed by the treatment manipulation

Post-Treatment Exclusion

- Any post-treatment exclusion is problematic and should be avoided
- Can estimate a LATE
  - Interpretation: Effect of manipulation check among those whose value of the check can be changed by the treatment manipulation
- Non-response or attrition is the same as researcher-imposed exclusion
  - Not problematic if MCAR
  - Nothing really to be done if caused by treatment



Introduction of "collider bias" wherein values of the manipulation check are affected by other factors.


Information
Opinion
Manipulation
Check
Etc.

Post-Treatment Exclusion

- Any post-treatment exclusion is problematic and should be avoided
- Can estimate a LATE
  - Interpretation: Effect of manipulation check among those whose value of the check can be changed by the treatment manipulation

Post-Treatment Exclusion

- Any post-treatment exclusion is problematic and should be avoided
- Can estimate a LATE
  - Interpretation: Effect of manipulation check among those whose value of the check can be changed by the treatment manipulation
- Non-response or attrition is the same as researcher-imposed exclusion
  - Not problematic if MCAR
  - Nothing really to be done if caused by treatment

Questions?

Treatments

- We should expect this! Why?

Treatments

- We should expect this! Why?
- What can we do?
  - Pilot testing
  - Replication
  - More complex design
  - Conjoint experiments

Conjoint Designs I

- "Classic vignettes" taken to an extreme
  - Address heterogeneity w/r/t SU**T**O
- Example: Judge whether to admit an immigrant to your country

Conjoint Designs I

- "Classic vignettes" taken to an extreme
  - Address heterogeneity w/r/t SU**T**O
- Example: Judge whether to admit an immigrant to your country
- Respondents see a series of vignettes that are fully randomized along any number of dimensions
  - Sex, Education, Language proficiency, etc.

Conjoint Designs I

- "Classic vignettes" taken to an extreme
  - Address heterogeneity w/r/t SU**T**O
- Example: Judge whether to admit an immigrant to your country
- Respondents see a series of vignettes that are fully randomized along any number of dimensions
  - Sex, Education, Language proficiency, etc.
- Outcome is judgment (binary or rating scale)

Conjoint Designs II

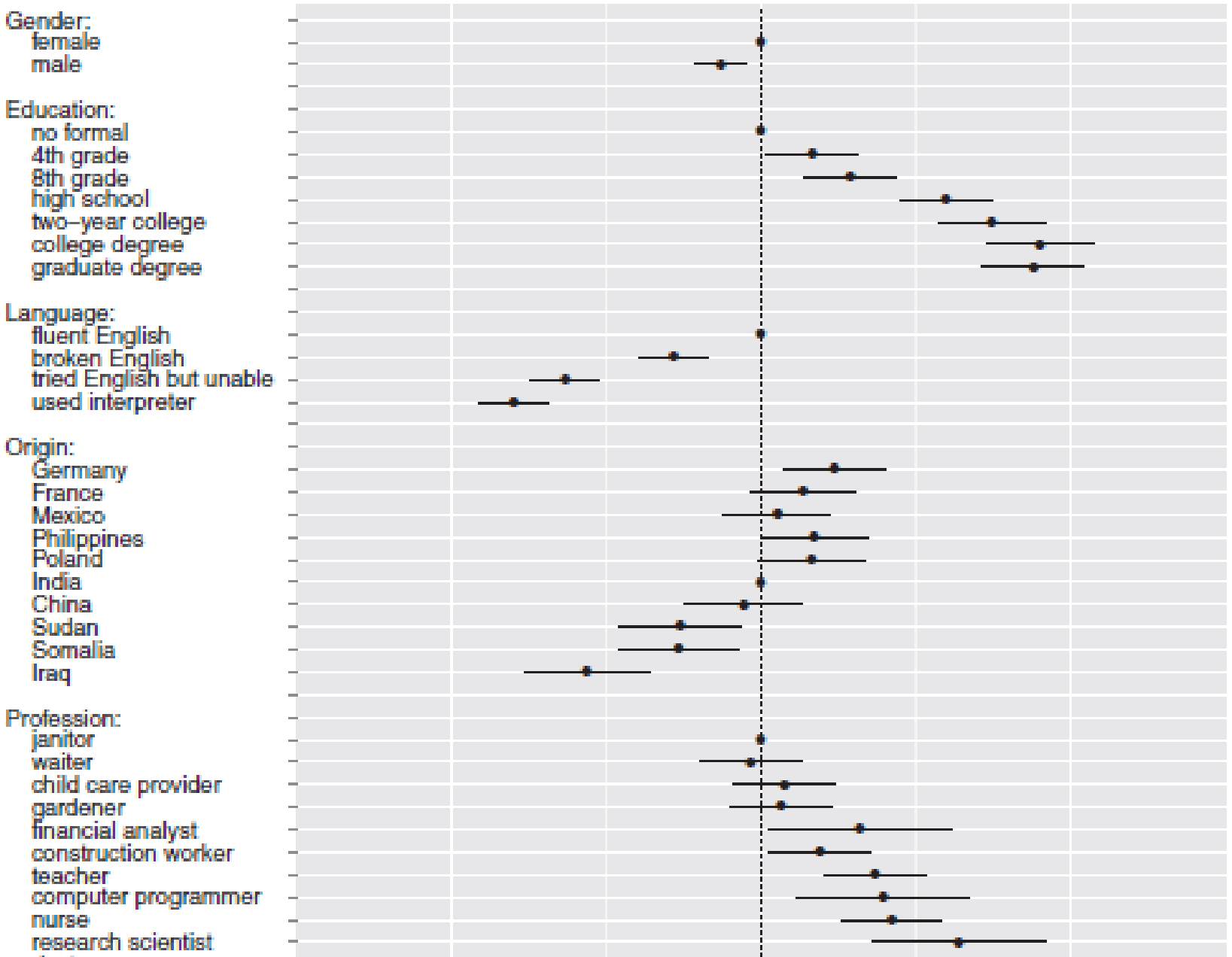
Why is this useful?

- Understand complex decision-making
- Within-subjects comparisons
- Heterogeneous effects across versions of treatment
- Pilot testing: Sensitivity of design to specification of *compound* vignette

Please read the descriptions of the potential immigrants carefully. Then, please indicate which of the two immigrants you would personally prefer to see admitted to the United States.

	Immigrant 1	Immigrant 2
Prior Trips to the U.S.	Entered the U.S. once before on a tourist visa	Entered the U.S. once before on a tourist visa
Reason for Application	Reunite with family members already in U.S.	Reunite with family members already in U.S.
Country of Origin	Mexico	Iraq
Language Skills	During admission interview, this applicant spoke fluent English	During admission interview, this applicant spoke fluent English
Profession	Child care provider	Teacher
Job Experience	One to two years of job training and experience	Three to five years of job training and experience
Employment Plans	Does not have a contract with a U.S. employer but has done job interviews	Will look for work after arriving in the U.S.
Education Level	Equivalent to completing two years of college in the U.S.	Equivalent to completing a college degree in the U.S.
Gender	Female	Male

	Immigrant 1	Immigrant 2
If you had to choose between them, which of these two immigrants should be given priority to come to the United States to live?	○	○


Gender:
female
male
Education:
no formal
4th grade
8th grade
high school
two−year college
college degree
graduate degree
Language:
fluent English
broken English
tried English but unable
used interpreter
Origin:
Germany
France
Mexico
Philippines
Poland
India
China
Sudan
Somalia
Iraq
Profession:
janitor
waiter
child care provider
gardener
financial analyst
construction worker
teacher
computer programmer
nurse
research scientist

Conjoint Designs III

Conjoint Designs III

- As long as profiles are randomized, this is just a complex factorial design where we can estimate *marginal effect* of each attribute
  - Treatment–control SATE, conditional on all other randomized factors

Conjoint Designs III

- As long as profiles are randomized, this is just a complex factorial design where we can estimate *marginal effect* of each attribute
  - Treatment–control SATE, conditional on all other randomized factors
- Assumptions:
  - Fully randomized profiles
  - No "carry-over" effects
  - No profile order effects

Replication

- Conjoints solve one problem: they identify the relative size of sources of heterogeneity within a given treatment

Replication

- Conjoints solve one problem: they identify the relative size of sources of heterogeneity within a given treatment
- But how should we consider experiments testing the same theory using different treatments?
  - "Triangulation"
  - Consistent directionality
  - Consistent (standardized) effect sizes

Replication

- Conjoints solve one problem: they identify the relative size of sources of heterogeneity within a given treatment
- But how should we consider experiments testing the same theory using different treatments?
  - "Triangulation"
  - Consistent directionality
  - Consistent (standardized) effect sizes
- Big conclusion: replication is important and there's not enough of it.

Questions?

Outcomes

- This is expected!
  - E.g., non-equivalent outcomes
- Reasonable to explore multiple outcomes
  - Multiple comparisons
  - Power considerations
  - Construct validity

Outcomes

- This is expected!
  - E.g., non-equivalent outcomes
- Reasonable to explore multiple outcomes
  - Multiple comparisons
  - Power considerations
  - Construct validity
- What outcomes you measure depend on your theory

Outcomes

- This is expected!
  - E.g., non-equivalent outcomes
- Reasonable to explore multiple outcomes
  - Multiple comparisons
  - Power considerations
  - Construct validity
- What outcomes you measure depend on your theory
- Lots of potential for behavioral measures!

Behavioural measures

Some behaviours that can be directly measured through survey questionnaires.

Behavioural measures

Some behaviours that can be directly measured through survey questionnaires.

Three broad categories:

Behavioural measures

Some behaviours that can be directly measured through survey questionnaires.

Three broad categories:

1. Behavioural measures that provide survey paradata

Behavioural measures

Some behaviours that can be directly measured through survey questionnaires.

Three broad categories:

1. Behavioural measures that provide survey paradata
2. Behavioural measures that operationalize attitudes

Behavioural measures

Some behaviours that can be directly measured through survey questionnaires.

Three broad categories:

1. Behavioural measures that provide survey paradata
2. Behavioural measures that operationalize attitudes
3. Behavioural measures that operationalize behaviours

Behavioural Measures for Paradata

Why?

- Respondents use of the survey tells us something meaningful about their behaviour

Behavioural Measures for Paradata

Why?

- Respondents use of the survey tells us something meaningful about their behaviour

What?

Behavioural Measures for Paradata

Why?

- Respondents use of the survey tells us something meaningful about their behaviour

What?

- Nonresponse

Behavioural Measures for Paradata

Why?

- Respondents use of the survey tells us something meaningful about their behaviour

What?

- Nonresponse
- Response latencies

Behavioural Measures for Paradata

Why?

- Respondents use of the survey tells us something meaningful about their behaviour

What?

- Nonresponse
- Response latencies
- Reading times

Behavioural Measures for Paradata

Why?

- Respondents use of the survey tells us something meaningful about their behaviour

What?

- Nonresponse
- Response latencies
- Reading times
- Answer switching

Behavioural Measures for Paradata

Why?

- Respondents use of the survey tells us something meaningful about their behaviour

What?

- Nonresponse
- Response latencies
- Reading times
- Answer switching
- Eye tracking

Behavioural Measures for Paradata

Why?

- Respondents use of the survey tells us something meaningful about their behaviour

What?

- Nonresponse
- Response latencies
- Reading times
- Answer switching
- Eye tracking
- Mouse tracking

Behavioural Measures for Paradata

Why?

- Respondents use of the survey tells us something meaningful about their behaviour

What?

- Nonresponse
- Response latencies
- Reading times
- Answer switching
- Eye tracking
- Mouse tracking
- Smartphone metadata

Behavioural Measures for Attitudes

Why?

- Attitudinal self-reports might be “cheap talk”

Behavioural Measures for Attitudes

Why?

- Attitudinal self-reports might be “cheap talk”

What?

Behavioural Measures for Attitudes

Why?

- Attitudinal self-reports might be “cheap talk”

What?

- Implicit Association Test

Behavioural Measures for Attitudes

Why?

- Attitudinal self-reports might be “cheap talk”

What?

- Implicit Association Test
- Incentivized Survey questions

Behavioural Measures for Behaviour

Why?

- We want to observe or affect behaviour (e.g., in an experiment)

Behavioural Measures for Behaviour

Why?

- We want to observe or affect behaviour (e.g., in an experiment)

What?

- Directly measure or initiate a direct measure of a behaviour
- May be measured by something that occurs within the confines of the survey or something outside of the survey

Example 1: Active Information Choice

[31] Guess, AM. 2015. "Measure for Measure." *Political Analysis* 23: 59–75. doi:10.1093/pan/mpu010

[32] Leeper, TJ. 2014. "The Informational Basis for Mass Polarization." *Public Opinion Quarterly* 78(1): 27–46. doi:10.1093/poq/nft045

[33] Arceneaux, K & Johnson, M. 2012. *Changing Minds or Changign Channels*. Chicago: The University of Chicago Press.

[34] `https://dpte.polisci.uiowa.edu/dpte/`

Example 1: Active Information Choice

- "Followed link" identification[31]

[31] Guess, AM. 2015. "Measure for Measure." *Political Analysis* 23: 59–75. doi:10.1093/pan/mpu010

[32] Leeper, TJ. 2014. "The Informational Basis for Mass Polarization." *Public Opinion Quarterly* 78(1): 27–46. doi:10.1093/poq/nft045

[33] Arceneaux, K & Johnson, M. 2012. *Changing Minds or Changign Channels*. Chicago: The University of Chicago Press.

[34] `https://dpte.polisci.uiowa.edu/dpte/`

Remember, please check **ALL** rows containing any links shown in **PURPLE**. Leave all other rows unchecked.

- [ ] LINK LINK
- [ ] LINK LINK
- [ ] LINK LINK
- [ ] LINK LINK
- [ ] LINK LINK LINK
- [ ] LINK LINK
- [ ] LINK LINK
- [ ] LINK LINK
- [ ] LINK LINK
- [ ] LINK LINK
- [ ] LINK
- [ ] LINK LINK LINK
- [ ] LINK
- [ ] LINK
- [ ] LINK LINK

Example 1: Active Information Choice

- "Followed link" identification[31]

[31] Guess, AM. 2015. "Measure for Measure." *Political Analysis* 23: 59–75. doi:10.1093/pan/mpu010

[32] Leeper, TJ. 2014. "The Informational Basis for Mass Polarization." *Public Opinion Quarterly* 78(1): 27–46. doi:10.1093/poq/nft045

[33] Arceneaux, K & Johnson, M. 2012. *Changing Minds or Changign Channels*. Chicago: The University of Chicago Press.

[34] `https://dpte.polisci.uiowa.edu/dpte/`

Example 1: Active Information Choice

- "Followed link" identification[31]
- Information boards[32]

[31] Guess, AM. 2015. "Measure for Measure." *Political Analysis* 23: 59–75. doi:10.1093/pan/mpu010

[32] Leeper, TJ. 2014. "The Informational Basis for Mass Polarization." *Public Opinion Quarterly* 78(1): 27–46. doi:10.1093/poq/nft045

[33] Arceneaux, K & Johnson, M. 2012. *Changing Minds or Changign Channels*. Chicago: The University of Chicago Press.

[34] `https://dpte.polisci.uiowa.edu/dpte/`

Reports From the Hive, Where the Swarm Concurs	Doctors Can Work Together to Improve Patient Health, But Need Appropriate Incentives	SEC Vote Requires Business Filings to Add Environmental Risks to Bottom Line	Wellness, Rather Than Illness, Is Focus Under Outcome-Accountable Care
Pay for Performance Improves Quality of Health Care Through Collaborative Medicine	Patients Better Served When Providers Paid for Health Outcomes	Anatomy of a Tear-Jerker	Gender Differences in Education Need Innovative Solution
Why are 3-D Movies so Bad?	Improving America's Health Requires Provider Incentives, Not 'Fee-for-Service'	Spammers Use the Human Touch to Avoid CAPTCHA	Heart Attack While Dining at Heart Attack Grill in Las Vegas
Physicians Group Says Quality Will Improve Under Outcome-based Payments	When Paid for Outcomes, Doctors Have Little Reason to Treat Highest Risk Patients	USDA Raises Corn Export Outlook	Out of the O.R., T.R. Knight Back Onto the Stage
Council Is Set to Consider Increases in Hotel and Property Taxes	A Bowl of Chili with Bragging Rights	Will a Standardized System for Verifying Web Identity Ever Catch On?	Paying Doctors Based on Outcomes Will Lead to Rationing

Example 1: Active Information Choice

- "Followed link" identification[31]
- Information boards[32]

[31] Guess, AM. 2015. "Measure for Measure." *Political Analysis* 23: 59–75. doi:10.1093/pan/mpu010

[32] Leeper, TJ. 2014. "The Informational Basis for Mass Polarization." *Public Opinion Quarterly* 78(1): 27–46. doi:10.1093/poq/nft045

[33] Arceneaux, K & Johnson, M. 2012. *Changing Minds or Changign Channels*. Chicago: The University of Chicago Press.

[34] `https://dpte.polisci.uiowa.edu/dpte/`

Example 1: Active Information Choice

- "Followed link" identification[31]
- Information boards[32]
- Video choice[33]

---

[31] Guess, AM. 2015. "Measure for Measure." *Political Analysis* 23: 59–75. doi:10.1093/pan/mpu010

[32] Leeper, TJ. 2014. "The Informational Basis for Mass Polarization." *Public Opinion Quarterly* 78(1): 27–46. doi:10.1093/poq/nft045

[33] Arceneaux, K & Johnson, M. 2012. *Changing Minds or Changign Channels*. Chicago: The University of Chicago Press.

[34] `https://dpte.polisci.uiowa.edu/dpte/`

Example 1: Active Information Choice

- "Followed link" identification[31]
- Information boards[32]
- Video choice[33]
- Dynamic Process Tracing Environment [34]

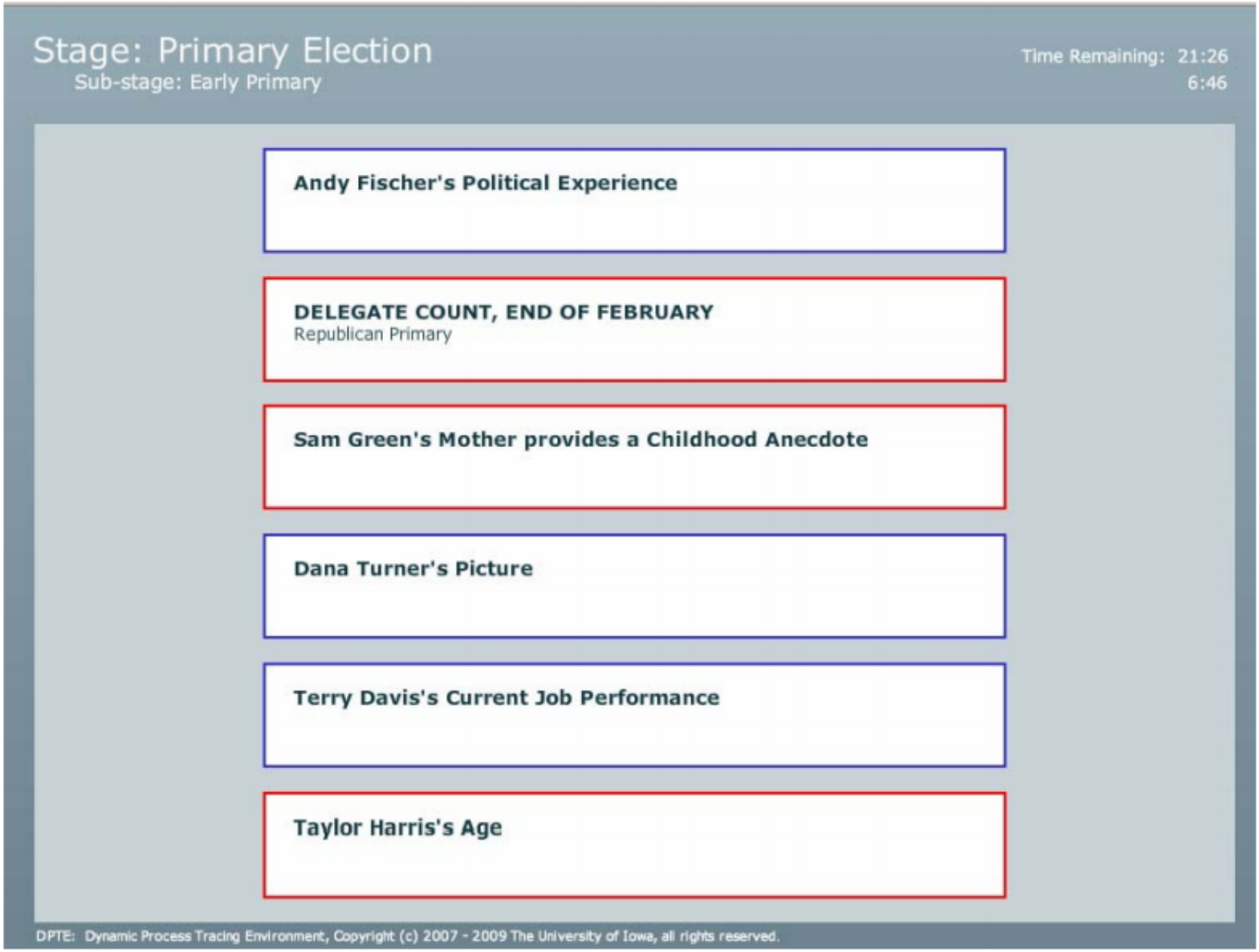
---

[31] Guess, AM. 2015. "Measure for Measure." *Political Analysis* 23: 59–75. doi:10.1093/pan/mpu010

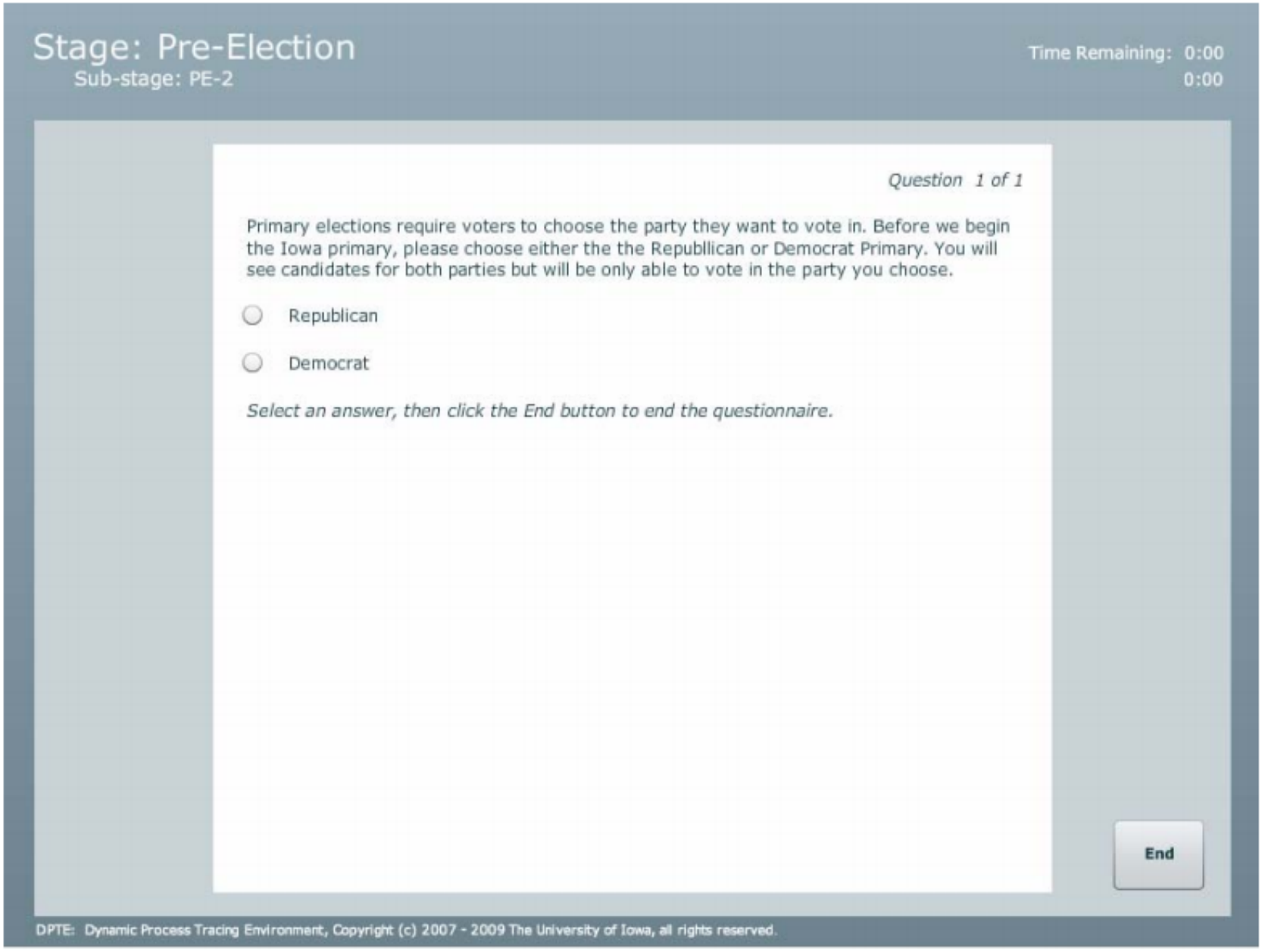
[32] Leeper, TJ. 2014. "The Informational Basis for Mass Polarization." *Public Opinion Quarterly* 78(1): 27–46. doi:10.1093/poq/nft045

[33] Arceneaux, K & Johnson, M. 2012. *Changing Minds or Changign Channels*. Chicago: The University of Chicago Press.

[34] `https://dpte.polisci.uiowa.edu/dpte/`


Stage: Primary Election
Sub-stage: Early Primary
Time Remaining: 21:26
6:46
Andy Fischer's Political Experience
DELEGATE COUNT, END OF FEBRUARY
Republican Primary
Sam Green's Mother provides a Childhood Anecdote
Dana Turner's Picture
Terry Davis's Current Job Performance
Taylor Harris's Age
DPTE: Dynamic Process Tracing Environment, Copyright (c) 2007 - 2009 The University of Iowa, all rights reserved.


Iowa General Election
January, 2008
Time remaining: 5:23
Hillary Clinton wins in South Dakota!
0:00 / 0:06
DPTE: Dynamic Process Tracing Environment, Copyright (c) 2007 - 2008 The University of Iowa, all rights reserved.


Stage: Pre-Election
Sub-stage: PE-2
Time Remaining: 0:00
0:00
Question 1 of 1
Primary elections require voters to choose the party they want to vote in. Before we begin the Iowa primary, please choose either the the Republlican or Democrat Primary. You will see candidates for both parties but will be only able to vote in the party you choose.
Republican
Democrat
Select an answer, then click the End button to end the questionnaire.
End
DPTE: Dynamic Process Tracing Environment, Copyright (c) 2007 - 2009 The University of Iowa, all rights reserved.

Example 2: Sign-up/Enrolment

An extension of information choice behaviour would be explicit engagement in other kinds of (small) behaviours, such as:

- Entering an email address to receive information or join a mailing list [35] [36]
- Signing up for an appointment or further interaction

[35] Leeper, TJ. 2017. "How Does Treatment Self-Selection Affect Inferences About Political Communication?" *Journal of Experimental Political Science*: In press.

[36] Bolsen, Druckman, & Cook. 2014. "Communication and Collective Actions." *Journal of Experimental Political Science* 1(1): 24–38. doi:10.1017/xps.2014.2

Example 3: Incentivised Survey Questions

Definitions:

- A survey question is just a self-report
- An *incentivized* survey question attached financial gains or losses to the answer options

Mark your gamble selection with an **X** in the last column across from your preferred gamble.

Gamble	Event	Payoff	Probabilities	Your Selection
1	A	$10	50%	
	B	$10	50%	
2	A	$18	50%	
	B	$6	50%	
3	A	$26	50%	
	B	$2	50%	
4	A	$34	50%	
	B	-$2	50%	
5	A	$42	50%	
	B	-$6	50%	

Eckel & Grossman. 2008 "Forecasting risk attitudes." *Journal of Economic Behavior & Organization* 68(1): 1–17. doi:10.1016/j.jebo.2008.04.006

Example 3: Incentivised Survey Questions

Definitions:

- A survey question is just a self-report
- An *incentivized* survey question attached financial gains or losses to the answer options

Example 3: Incentivised Survey Questions

Definitions:

- A survey question is just a self-report
- An *incentivized* survey question attached financial gains or losses to the answer options

Paradigm could be applied to any measure of behavioural intentions to avoid cheap talk.

Example 4: Purchasing Decisions

Common ways to study purchasing behaviour include:

Example 4: Purchasing Decisions

Common ways to study purchasing behaviour include:

- Direct attitudinal questions

Example 4: Purchasing Decisions

Common ways to study purchasing behaviour include:

- Direct attitudinal questions
- Retrospective and prospective self-reports

Example 4: Purchasing Decisions

Common ways to study purchasing behaviour include:

- Direct attitudinal questions
- Retrospective and prospective self-reports
- Conjoint experiments

Example 4: Purchasing Decisions

Common ways to study purchasing behaviour include:

- Direct attitudinal questions
- Retrospective and prospective self-reports
- Conjoint experiments

Another way is embedding a purchase in a survey.[37]

[37] Bolsen, T. 2011. "A Lightbulb Goes On." *Political Behavior* 35(1): 1–20. 10.1007/s11109-011-9186-5

Source: Wikimedia Commons (Sun Ladder, KMJ)

Example 5: Donations

- Miller and Krosnick[38] asked for charitable donations via cheque directly as part of a paper-and-pencil survey

[38] Miller, Krosnick, & Lowe. N.d. "The Impact of Policy Change Threat on Financial Contributions to Interest Groups." Working paper.

[39] Klar & Piston. 2015. "The influence of competing organisational appeals on individual donations." *Journal of Public Policy* 35(2): 171–91. doi:10.1017/S0143814X15000203

Example 5: Donations

- Miller and Krosnick[38] asked for charitable donations via cheque directly as part of a paper-and-pencil survey
- Klar and Piston[39] offered respondents a survey incentive up-front for participation and then later offered them a chance to donate (a portion of payment) to a charity

[38] Miller, Krosnick, & Lowe. N.d. "The Impact of Policy Change Threat on Financial Contributions to Interest Groups." Working paper.

[39] Klar & Piston. 2015. "The influence of competing organisational appeals on individual donations." *Journal of Public Policy* 35(2): 171–91. doi:10.1017/S0143814X15000203

Example 6: Web Tracking Data

1. Active installation of a tracking app, such as YouGov Pulse[40] [41]
2. Post-hoc collection of web history files using something like Web Historian [42]

---

[40] https://yougov.co.uk/find-solutions/profiles/pulse/

[41] Guess, AM. N.d. "Media Choice and Moderation." Working paper, https://dl.dropboxusercontent.com/u/663930/GuessJMP.pdf.

[42] http://www.webhistorian.org/

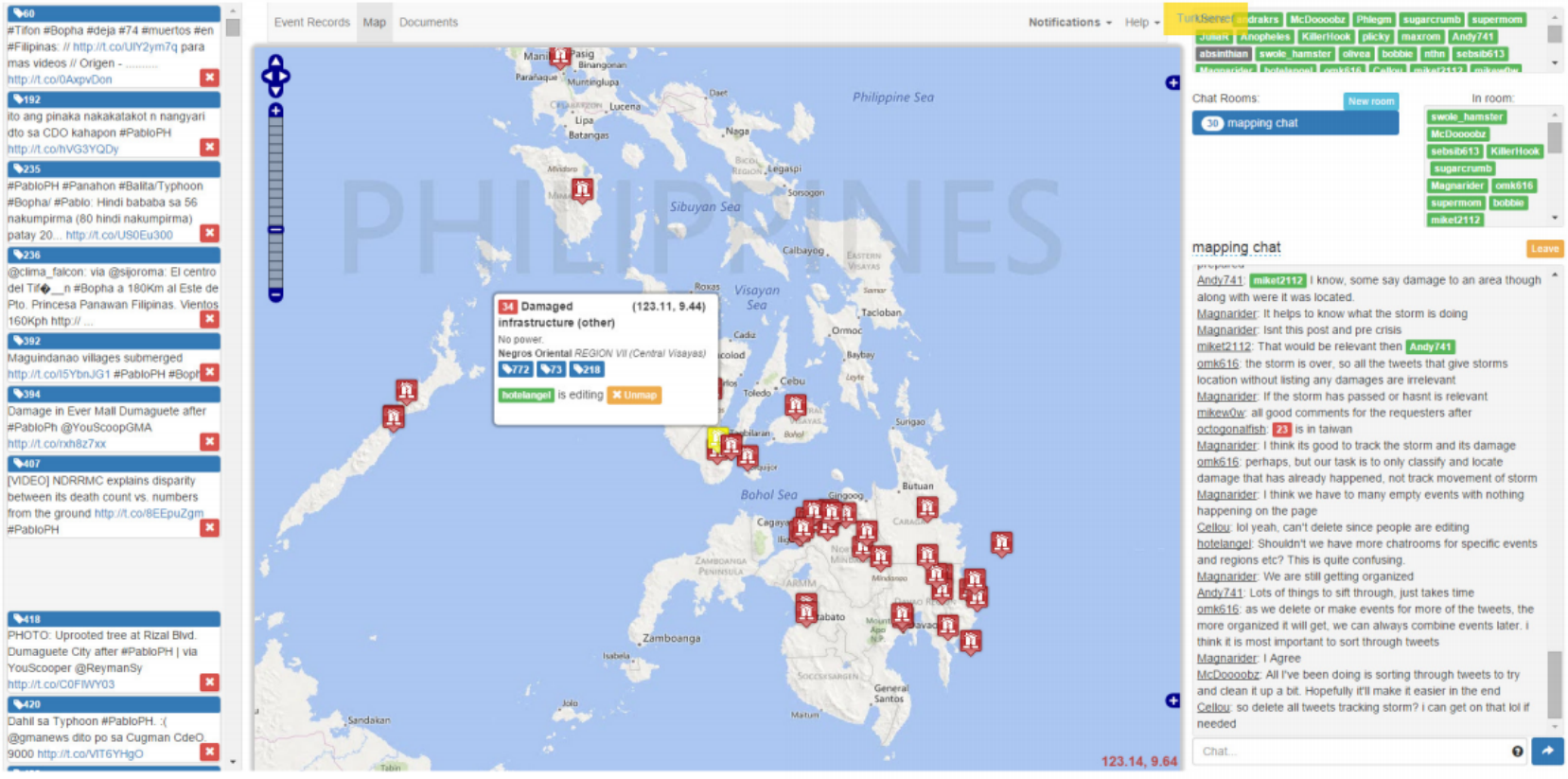
Other Possibilities

[43] Mao, Mason, Suri, Watts. 2016. "An Experimental Study of Team Size and Performance on a Complex Task." *PLoS ONE* 11(4): e0153048. doi:10.1371/journal.pone.0153048

Other Possibilities

- Coordination tasks
  - Synchronous group tasks[43]
  - Game play
  - Simulations

[43] Mao, Mason, Suri, Watts. 2016. "An Experimental Study of Team Size and Performance on a Complex Task." *PLoS ONE* 11(4): e0153048. doi:10.1371/journal.pone.0153048


#60
#Tifon #Bopha #deja #74 #muertos #en #Filipinas: // http://t.co/UlY2ym7q para mas videos // Origen - .....
http://t.co/0AxpvDon
#192
ito ang pinaka nakakatakot n nangyari dto sa CDO kahapon #PabloPH
http://t.co/hVG3YQDy
#235
#PabloPH #Panahon #Balita/Typhoon #Bopha/ #Pablo: Hindi bababa sa 56 nakumpirma (80 hindi nakumpirma) patay 20... http://t.co/US0Eu300
#236
@clima_falcon: via @sijoroma: El centro del Tif_n #Bopha a 180Km al Este de Pto. Princesa Panawan Filipinas. Vientos 160Kph http:// ...
#392
Maguindanao villages submerged
http://t.co/I5YbnJG1 #PabloPH #Bopha
#394
Damage in Ever Mall Dumaguete after #PabloPh @YouScoopGMA
http://t.co/rxh8z7xx
#407
[VIDEO] NDRRMC explains disparity between its death count vs. numbers from the ground http://t.co/8EEpuZgm #PabloPH
#418
PHOTO: Uprooted tree at Rizal Blvd. Dumaguete City after #PabloPH | via YouScooper @ReymanSy
http://t.co/C0FIWY03
#420
Dahil sa Typhoon #PabloPH. :(@gmanews dito po sa Cugman CdeO... 9000 http://t.co/VlT6YHgO
Event Records
Map
Documents
Notifications
Help
Philippine Sea
Sibuyan Sea
Visayan Sea
Bohol Sea
Zamboanga
34 Damaged infrastructure (other)
(123.11, 9.44)
No power.
Negros Oriental REGION VII (Central Visayas)
772
73
218
hotelangel is editing
Unmap
123.14, 9.64
Chat Rooms:
New room
30 mapping chat
In room:
swole_hamster
McDoooobz
sebsib613
KillerHook
sugarcrumb
Magnarider
omk616
supermom
bobbie
miket2112
mapping chat
Leave
Andy741: miket2112 I know, some say damage to an area though along with were it was located.
Magnarider: It helps to know what the storm is doing
Magnarider: Isnt this post and pre crisis
miket2112: That would be relevant then Andy741
omk616: the storm is over, so all the tweets that give storms location without listing any damages are irrelevant
Magnarider: If the storm has passed or hasnt is relevant
mikew0w: all good comments for the requesters after
octogonalfish: 23 is in taiwan
Magnarider: I think its good to track the storm and its damage
omk616: perhaps, but our task is to only classify and locate damage that has already happened, not track movement of storm
Magnarider: I think we have to many empty events with nothing happening on the page
Cellou: lol yeah, can't delete since people are editing
hotelangel: Shouldn't we have more chatrooms for specific events and regions etc? This is quite confusing.
Magnarider: We are still getting organized
Andy741: Lots of things to sift through, just takes time
omk616: as we delete or make events for more of the tweets, the more organized it will get, we can always combine events later. i think it is most important to sort through tweets
Magnarider: I Agree
McDoooobz: All I've been doing is sorting through tweets to try and clean it up a bit. Hopefully it'll make it easier in the end
Cellou: so delete all tweets tracking storm? i can get on that lol if needed
Chat...

Other Possibilities

- Coordination tasks
  - Synchronous group tasks[43]
  - Game play
  - Simulations

[43] Mao, Mason, Suri, Watts. 2016. "An Experimental Study of Team Size and Performance on a Complex Task." *PLoS ONE* 11(4): e0153048. doi:10.1371/journal.pone.0153048

Other Possibilities

- Coordination tasks
  - Synchronous group tasks[43]
  - Game play
  - Simulations

- Offering incentives to perform future behaviour (tracked elsewhere)

[43] Mao, Mason, Suri, Watts. 2016. "An Experimental Study of Team Size and Performance on a Complex Task." *PLoS ONE* 11(4): e0153048. doi:10.1371/journal.pone.0153048

Other Possibilities

- Coordination tasks
  - Synchronous group tasks[43]
  - Game play
  - Simulations
- Offering incentives to perform future behaviour (tracked elsewhere)
- OAuth/API integrations w/ other platforms
  - Merging website usage data w/ survey data
  - Treating website sign-up or usage as behavioural outcomes
  - Linking with smartphone metadata

[43] Mao, Mason, Suri, Watts. 2016. "An Experimental Study of Team Size and Performance on a Complex Task." *PLoS ONE* 11(4): e0153048. doi:10.1371/journal.pone.0153048

Activity!

With a partner, brainstorm how one or more these behavioural measures might be applied to a survey data collection relevant to your own work or your organisation.

"SUTO" Punchline: Replication!

- If we think effects are homogeneous (across SUTO), then replications in other SUTO conditions should provide us the same SATE (within sampling error)
- If we think effects are heterogeneous, then replications should give *systematically* different SATE (or CATE) estimates

"SUTO" Punchline: Replication!

- If we think effects are homogeneous (across SUTO), then replications in other SUTO conditions should provide us the same SATE (within sampling error)
- If we think effects are heterogeneous, then replications should give *systematically* different SATE (or CATE) estimates
  - Identify those patterns of heterogeneity using meta-analysis

"SUTO" Punchline: Replication!

- If we think effects are homogeneous (across SUTO), then replications in other SUTO conditions should provide us the same SATE (within sampling error)
- If we think effects are heterogeneous, then replications should give *systematically* different SATE (or CATE) estimates
  - Identify those patterns of heterogeneity using meta-analysis
  - Regress effect estimates from multiple studies on SUTO features of each study

Conclusion

- Do we want to know SATE, CATE(s), or both?

Conclusion

- Do we want to know SATE, CATE(s), or both?
- Decide in advance
  - Include in protocol
  - Design study to estimate CATE(s)

Conclusion

- Do we want to know SATE, CATE(s), or both?
- Decide in advance
  - Include in protocol
  - Design study to estimate CATE(s)
- Estimation of unit-related CATEs
  - Block randomization
  - Post-hoc procedures

Questions?

Beyond One-shot Designs

- Surveys can be used as a measurement instrument for a field treatment or a manipulation applied in a different survey panel wave
  1. Measure effect duration in two-wave panel
  2. Solicit pre-treatment outcome measures in a two-wave panel
  3. Measure effects of field treatment in post-test only design
  4. Randomly encourage field treatment in pre-test and measure effects in post-test

Beyond One-shot Designs

- Surveys can be used as a measurement instrument for a field treatment or a manipulation applied in a different survey panel wave
  1. Measure effect duration in two-wave panel
  2. Solicit pre-treatment outcome measures in a two-wave panel
  3. Measure effects of field treatment in post-test only design
  4. Randomly encourage field treatment in pre-test and measure effects in post-test
- Problems? Compliance & nonresponse

I. Effect Duration

- Use a two- (or more-) wave panel to measure duration of effects
  - T1: Treatment and outcome measurement
  - T2+: Outcome measurement
- Two main concerns
  - Attrition
  - Panel conditioning

II. Within-Subjects Designs

- Estimate treatment effects as a difference-in-differences
- Instead of using the post-treatment mean-difference in Y to estimate the causal effect, use the difference in pre-post differences for the two groups:

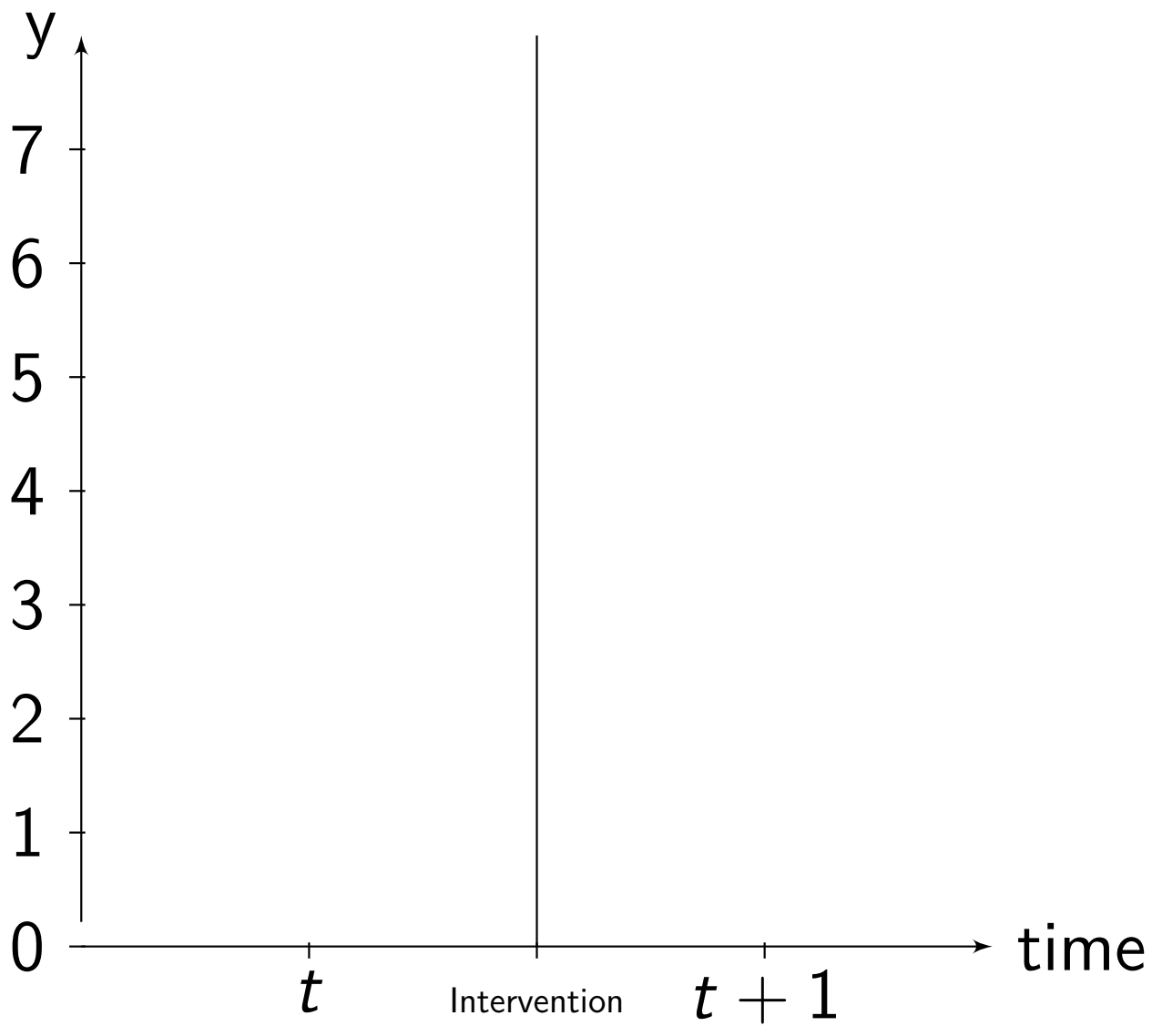
$$(\hat{Y}_{0,t+1} - \hat{Y}_{0,t}) - (\hat{Y}_{j,t+1} - \hat{Y}_{j,t})$$

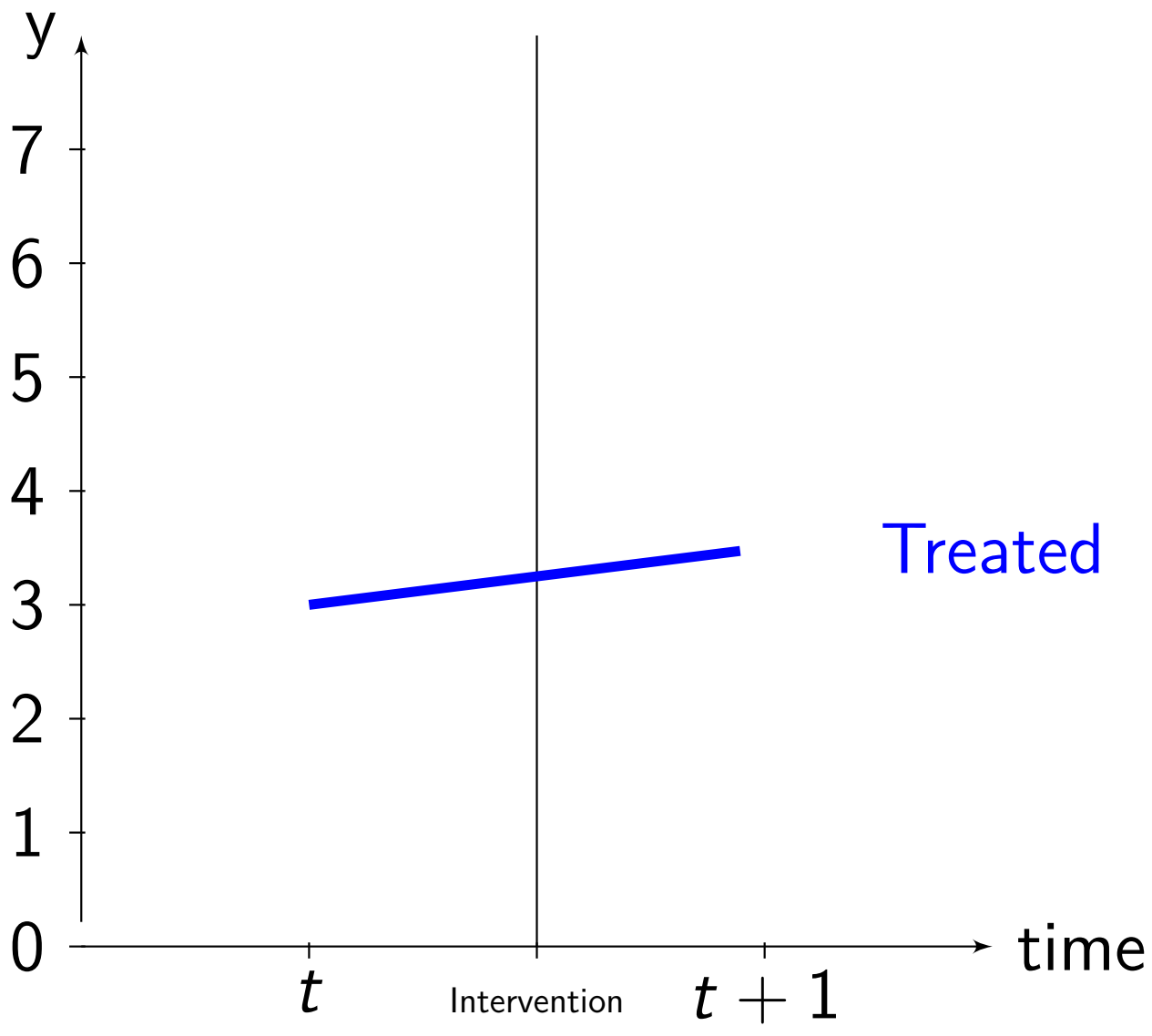
II. Within-Subjects Designs

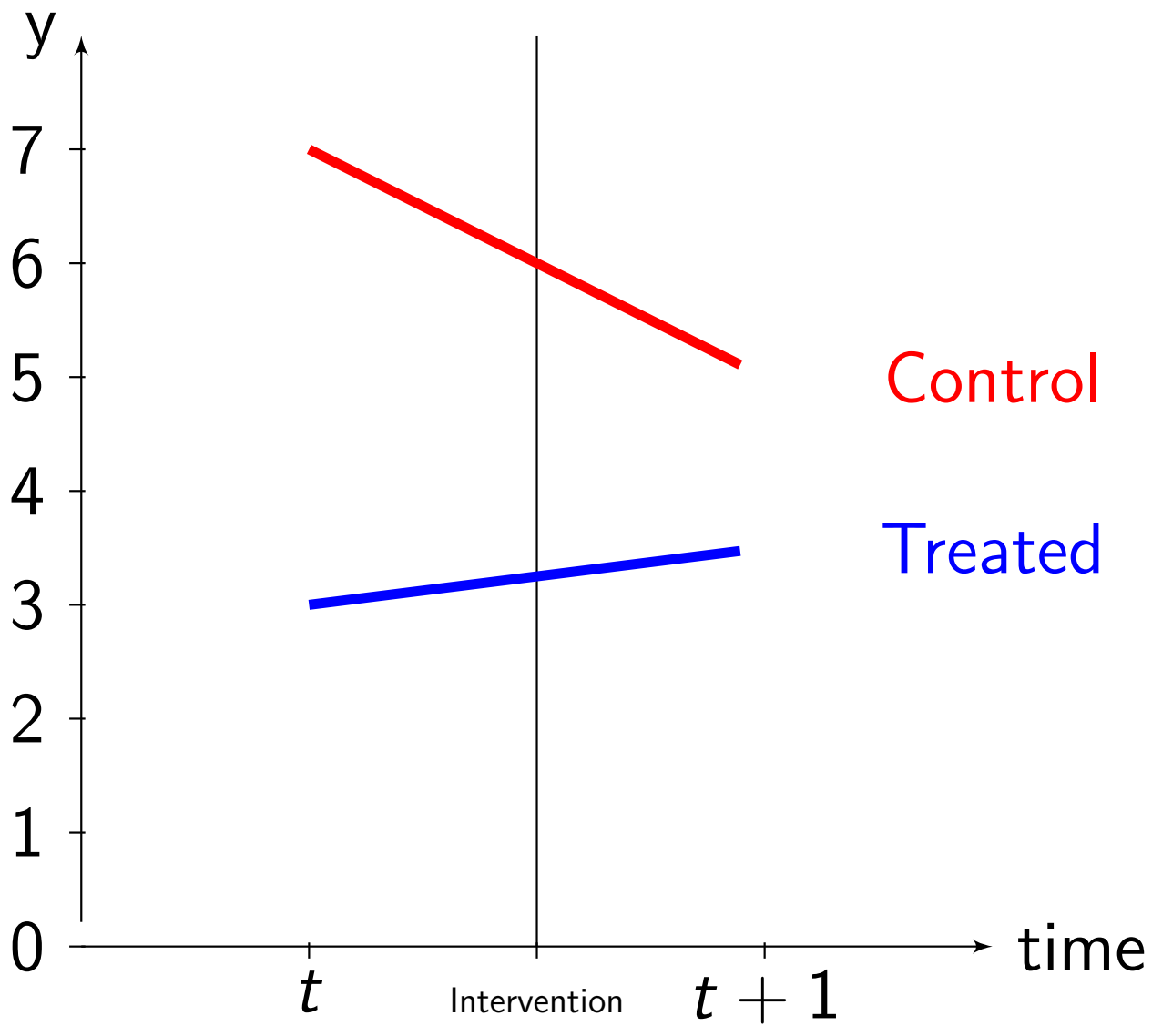
- Estimate treatment effects as a difference-in-differences
- Instead of using the post-treatment mean-difference in Y to estimate the causal effect, use the difference in pre-post differences for the two groups:

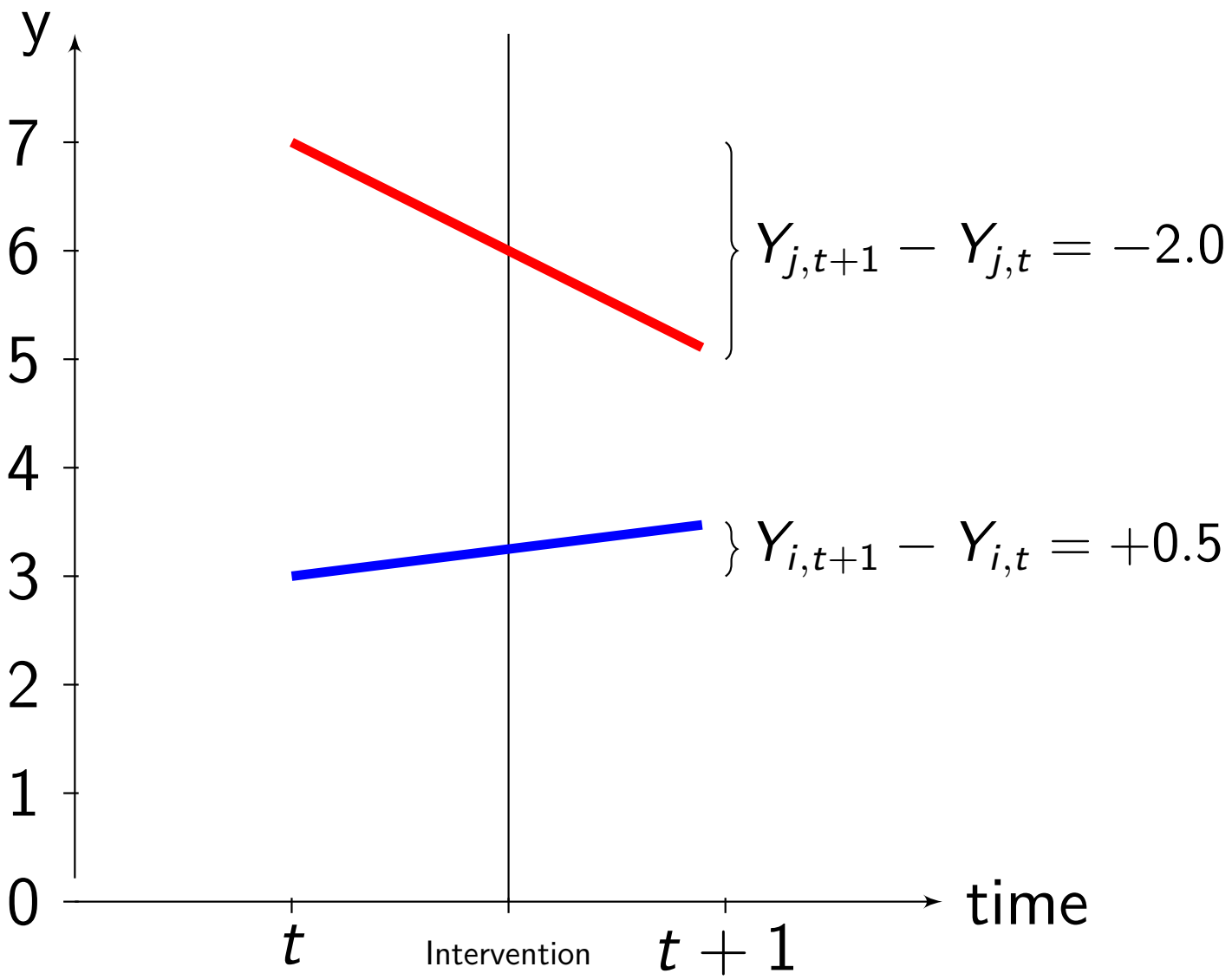
$$(\hat{Y}_{0,t+1} - \hat{Y}_{0,t}) - (\hat{Y}_{j,t+1} - \hat{Y}_{j,t})$$

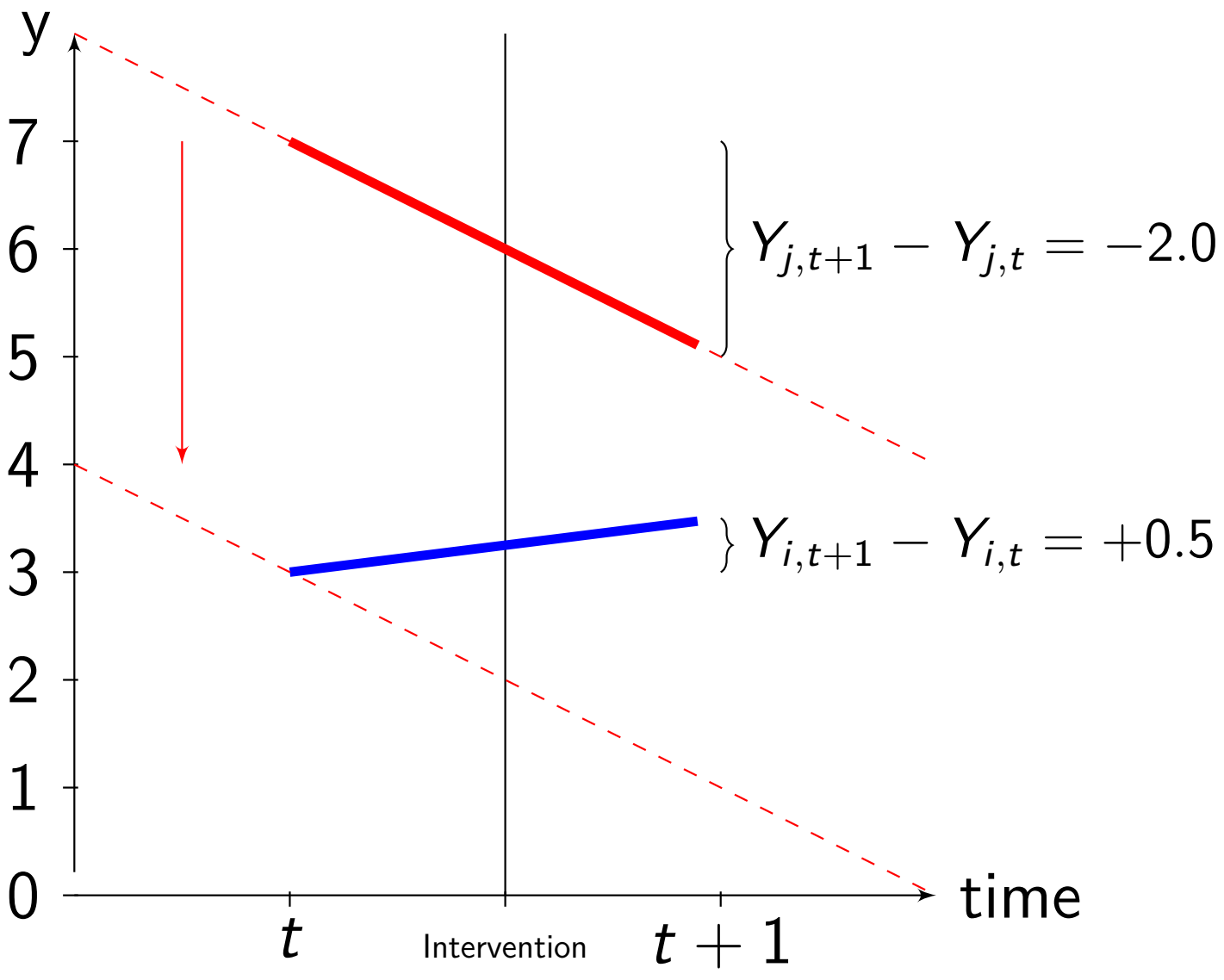
- Advantageous because variance for paired samples decreases as correlation between t_0 and t_1 observations increases

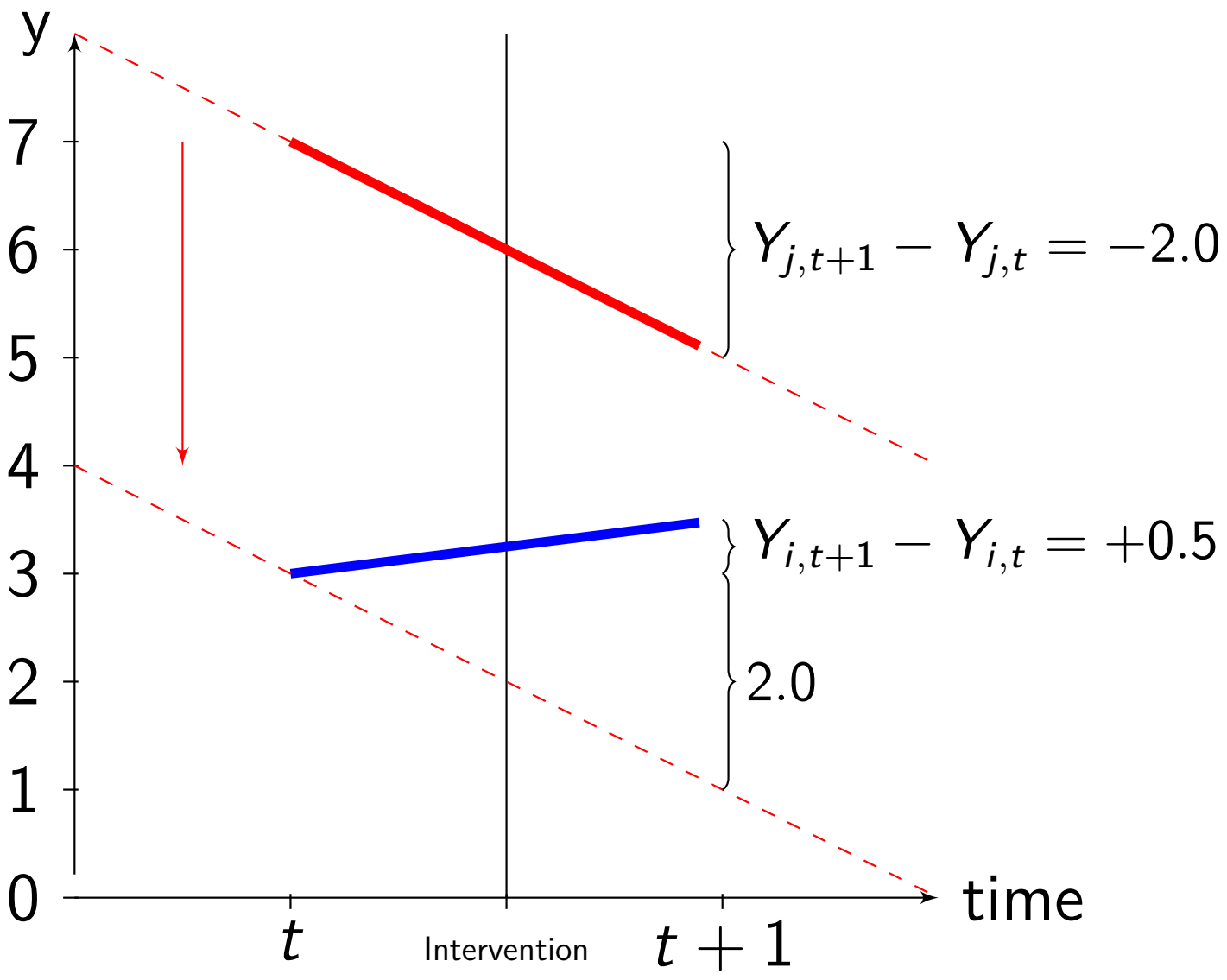

y
7
6
5
4
3
2
1
0
time
t
Intervention
$t+1$

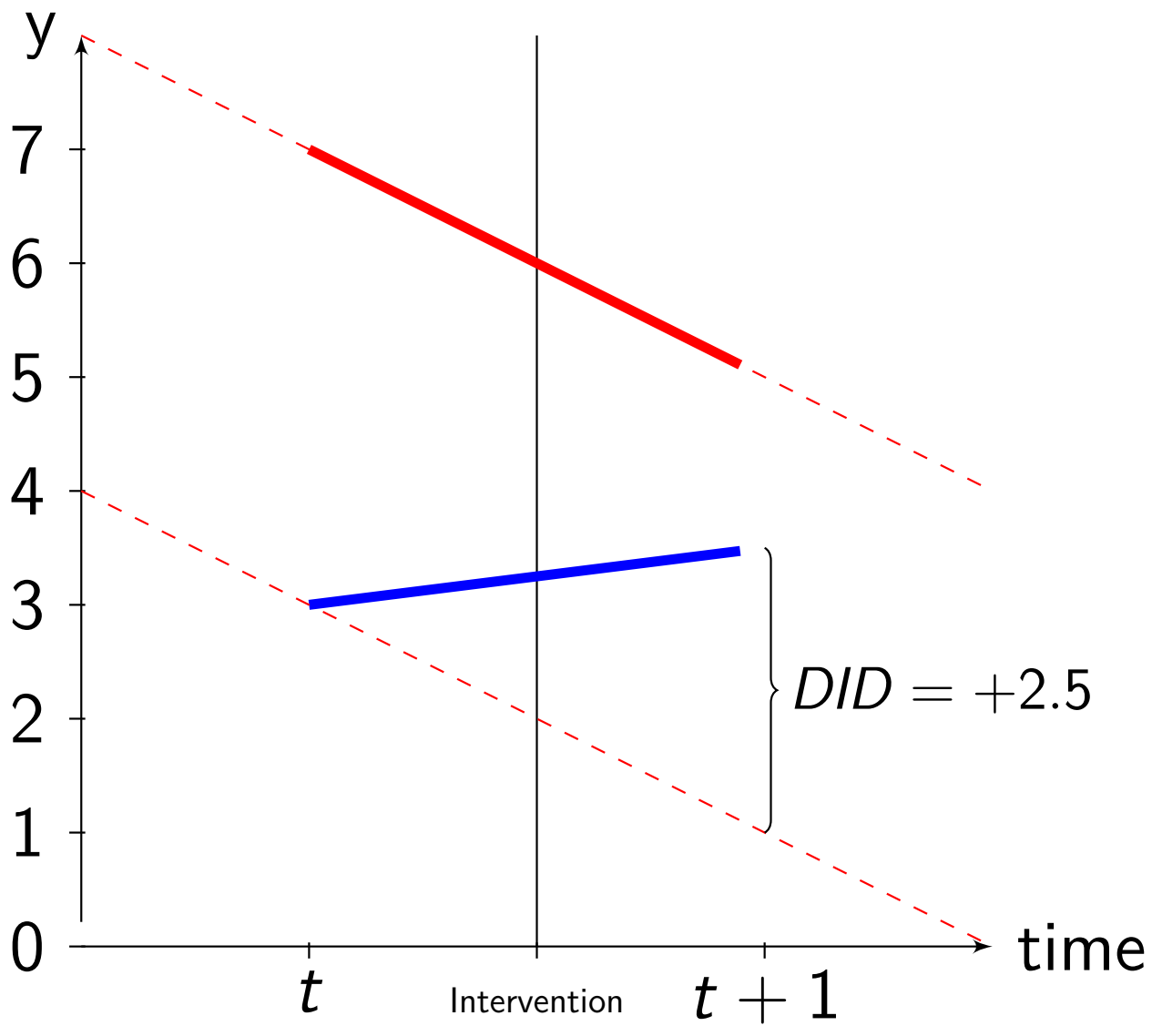

y
7
6
5
4
3
2
1
0
Treated
time
t
Intervention
$t+1$


y
7
6
5
4
3
2
1
0
Control
Treated
time
t
Intervention
$t+1$


y
7
6
5
4
3
2
1
0
$Y_{j,t+1} - Y_{j,t} = -2.0$
$Y_{i,t+1} - Y_{i,t} = +0.5$
time
t
Intervention
$t+1$


y
7
6
5
4
3
2
1
0
$Y_{j,t+1} - Y_{j,t} = -2.0$
$Y_{i,t+1} - Y_{i,t} = +0.5$
time
t
Intervention
$t+1$


y
7
6
5
4
3
2
1
0
$Y_{j,t+1} - Y_{j,t} = -2.0$
$Y_{i,t+1} - Y_{i,t} = +0.5$
2.0
time
t
Intervention
$t+1$


y
7
6
5
4
3
2
1
0
time
t
Intervention
$t+1$
$DID = +2.5$

Threats to Validity

As soon as time comes into play, we have to worry about threats to validity.[44]

[44] Shadish, Cook, and Campbell (2002)

Threats to Validity

As soon as time comes into play, we have to worry about threats to validity.[44]

1. History (simultaneous cause)

[44] Shadish, Cook, and Campbell (2002)

Threats to Validity

As soon as time comes into play, we have to worry about threats to validity.[44]

1. History (simultaneous cause)
2. Maturation (time trends)

[44] Shadish, Cook, and Campbell (2002)

Threats to Validity

As soon as time comes into play, we have to worry about threats to validity.[44]

1. History (simultaneous cause)
2. Maturation (time trends)
3. Testing (observation changes respondents)

[44] Shadish, Cook, and Campbell (2002)

Threats to Validity

As soon as time comes into play, we have to worry about threats to validity.[44]

1. History (simultaneous cause)
2. Maturation (time trends)
3. Testing (observation changes respondents)
4. Instrumentation (changing operationalization)

[44] Shadish, Cook, and Campbell (2002)

Threats to Validity

As soon as time comes into play, we have to worry about threats to validity.[44]

1. History (simultaneous cause)
2. Maturation (time trends)
3. Testing (observation changes respondents)
4. Instrumentation (changing operationalization)
5. Instability (measurement error)

[44] Shadish, Cook, and Campbell (2002)

Threats to Validity

As soon as time comes into play, we have to worry about threats to validity.[44]

1. History (simultaneous cause)
2. Maturation (time trends)
3. Testing (observation changes respondents)
4. Instrumentation (changing operationalization)
5. Instability (measurement error)
6. Attrition

[44] Shadish, Cook, and Campbell (2002)

III. Randomized Field Treatment

- Examples:

III. Randomized Field Treatment

- Examples:
  1. Citizens randomly sent a letter by post encouraging them to reduce water usage

III. Randomized Field Treatment

- Examples:
  1. Citizens randomly sent a letter by post encouraging them to reduce water usage
  2. Different local media markets randomly assigned to receive different advertising

III. Randomized Field Treatment

- Examples:
  1. Citizens randomly sent a letter by post encouraging them to reduce water usage
  2. Different local media markets randomly assigned to receive different advertising
- Survey is used to measure outcomes, when treatment assignment is already known

III. Randomized Field Treatment

- Examples:
  1. Citizens randomly sent a letter by post encouraging them to reduce water usage
  2. Different local media markets randomly assigned to receive different advertising
- Survey is used to measure outcomes, when treatment assignment is already known
- Issues

III. Randomized Field Treatment

- Examples:
  1. Citizens randomly sent a letter by post encouraging them to reduce water usage
  2. Different local media markets randomly assigned to receive different advertising
- Survey is used to measure outcomes, when treatment assignment is already known
- Issues
  - Nonresponse
  - Noncompliance

IV. Treatment Encouragement

- Design:
  - T1: Encourage treatment
  - T2: Measure effects
- Examples:
  1. Albertson and Lawrence[45]

[45] Albertson & Lawrence. 2009. "After the Credits Roll." *American Politics Research* 37(2): 275–300. 10.1177/1532673X08328600.

IV. Treatment Encouragement

- Design:
  - T1: Encourage treatment
  - T2: Measure effects
- Examples:
  1. Albertson and Lawrence[45]
- Issues

[45] Albertson & Lawrence. 2009. "After the Credits Roll." *American Politics Research* 37(2): 275–300. 10.1177/1532673X08328600.

IV. Treatment Encouragement

- Design:
  - T1: Encourage treatment
  - T2: Measure effects
- Examples:
  1. Albertson and Lawrence[45]
- Issues
  - Nonresponse
  - Noncompliance

[45] Albertson & Lawrence. 2009. "After the Credits Roll." *American Politics Research* 37(2): 275–300. 10.1177/1532673X08328600.

Treatment Noncompliance

- Definition:

  "when subjects who were assigned to receive the treatment go untreated or when subjects assigned to the control group are treated" [46]

[46] Gerber & Green. 2012. *Field Experiments*, p.132.

Treatment Noncompliance

- Definition:

  "when subjects who were assigned to receive the treatment go untreated or when subjects assigned to the control group are treated" [46]

- Several strategies
  - "As treated" analysis
  - "Intention to treat" analysis
  - Estimate a LATE

[46] Gerber & Green. 2012. *Field Experiments*, p.132.

Analyzing Noncompliance

- If noncompliance only occurs in one group, it is *asymmetric* or *one-sided*
- We can ignore non-compliance and analyze the "intention to treat" effect, which will underestimate our effects because some people were not treated as assigned: $ITT = \overline{Y}_1 - \overline{Y}_0$

Analyzing Noncompliance

- If noncompliance only occurs in one group, it is *asymmetric* or *one-sided*
- We can ignore non-compliance and analyze the "intention to treat" effect, which will underestimate our effects because some people were not treated as assigned: $ITT = \overline{Y}_1 - \overline{Y}_0$
- We can use "instrumental variables" to estimate the "local average treatment effect" (LATE) for those that complied with treatment: $LATE = \frac{ITT}{\%Compliant}$

Local Average Treatment Effect

- IV estimate is *local* to the variation in X that is due to variation in D
- This matters if effects are *heterogeneous*
- LATE is effect for those who *comply*
- Four subpopulations:
  - Compliers: $X = 1$ only if $D = 1$
  - Always-takers: $X = 1$ regardless of D
  - Never-takers: $X = 0$ regardless of D
  - Defiers: $X = 1$ only if $D = 0$
- Exclusion restriction! Monotonicity!

Questions?

Quiz time!

Compliance

1. What is compliance?

Compliance

1. What is compliance?
2. How can we analyze experimental data when there is noncompliance?

Balance testing

1. What does randomization ensure about the composition of treatment groups?

Balance testing

1. What does randomization ensure about the composition of treatment groups?
2. What can we do if we find a covariate imbalance between groups?

Balance testing

1. What does randomization ensure about the composition of treatment groups?
2. What can we do if we find a covariate imbalance between groups?
3. How can we avoid this problem entirely?

Nonresponse and Attrition

1. Do we care about outcome nonresponse in experiments?

Nonresponse and Attrition

1. Do we care about outcome nonresponse in experiments?
2. How can we analyze experimental data when there is outcome nonresponse or post-treatment attrition?

Manipulation checks

1. What is a manipulation check? What can we do with it?

Manipulation checks

1. What is a manipulation check? What can we do with it?
2. What do we do if some respondents "fail" a manipulation check?

Null effects

1. What should we do if we find our estimated $\widehat{SATE} = 0$?

Null effects

1. What should we do if we find our estimated $\widehat{SATE} = 0$?
2. What does it mean for an experiment to be *underpowered*?

Null effects

1. What should we do if we find our estimated $\widehat{SATE} = 0$?
2. What does it mean for an experiment to be *underpowered*?
3. What can we do to reduce the probability of obtaining an (unwanted) "null effect"?

Effect heterogeneity

1. What should we do if, post-hoc, we find evidence of effect heterogeneity?

Effect heterogeneity

1. What should we do if, post-hoc, we find evidence of effect heterogeneity?
2. What can we do pre-implementation to address possible heterogeneity?

Representativeness

1. Under what conditions is a design-based, probability sample necessary for experimental inference?

Representativeness

1. Under what conditions is a design-based, probability sample necessary for experimental inference?
2. What kind of causal inferences can we draw from an experiment on a descriptively unrepresentative sample?

Peer Review

1. What should we do if a peer reviewer asks us to “control” for covariates in the analysis?

Peer Review

1. What should we do if a peer reviewer asks us to "control" for covariates in the analysis?
2. What should we do if a peer reviewer asks us to include or exclude particular respondents from the analysis?

Questions?

Presentations!

Look for TESS Examples

In groups of 2–3, look through some TESS examples

- What was the researcher's question?
- How did they test it experimentally?
- What was interesting or surprising about the designs?

Take about 15 minutes.

Learning Outcomes

By the end of the day, you should be able to...

Learning Outcomes

By the end of the day, you should be able to...

1. Explain how to analyze experiments quantitatively.

Learning Outcomes

By the end of the day, you should be able to...

1. Explain how to analyze experiments quantitatively.
2. Explain how to design experiments that speak to relevant research questions and theories.

Learning Outcomes

By the end of the day, you should be able to...

1. Explain how to analyze experiments quantitatively.
2. Explain how to design experiments that speak to relevant research questions and theories.
3. Evaluate the uses and limitations of several common survey experimental paradigms.

Learning Outcomes

By the end of the day, you should be able to...

1. Explain how to analyze experiments quantitatively.
2. Explain how to design experiments that speak to relevant research questions and theories.
3. Evaluate the uses and limitations of several common survey experimental paradigms.
4. Identify practical issues that arise in the implementation of experiments and evaluate how to anticipate and respond to them.

Wrap-up

- Thanks to all of you!
- Stay in touch (t.leeper@lse.ac.uk)
- Good luck with your research!

Apparent Satisficing

- Some common measures:
  - "Straightlining"
  - Non-differentiation
  - Acquiescence
  - Nonresponse
  - DK responding
  - Speeding
- Difficult to detect and distinguish from "real" responses

Metadata/Paradata

- Timing
  - Some survey tools will allow you to time page
  - Make a prior rules about dropping participants for speeding

Metadata/Paradata

- Timing
  - Some survey tools will allow you to time page
  - Make a prior rules about dropping participants for speeding
- Mousetracking or eyetracking
  - Mousetracking is unobtrusive
  - Eyetracking requires participants opt-in

Metadata/Paradata

- Timing
  - Some survey tools will allow you to time page
  - Make a prior rules about dropping participants for speeding
- Mousetracking or eyetracking
  - Mousetracking is unobtrusive
  - Eyetracking requires participants opt-in
- Record focus/blur browser events

Direct Measures

- How closely have you been paying attention to what the questions on this survey actually mean?

Direct Measures

- How closely have you been paying attention to what the questions on this survey actually mean?
- While taking this survey, did you engage in any of the following behaviors? Please check all that apply.
  - Use your mobile phone
  - Browse the internet
  - ...

Instructional Manipulation Check

We would like to know if you are reading the questions on this survey. If you are reading carefully, please ignore this question, do not select any answer below, and click "next" to proceed with the survey.

Strongly disagree
Somewhat disagree
Neither agree nor disagree
Somewhat agree
Strongly agree

Instructional Manipulation Check

Do you agree or disagree with the decision to send British forces to fight ISIL in Syria? We would like to know if you are reading the questions on this survey. If you are reading carefully, please ignore this question, do not select any answer below, and click "next" to proceed with the survey.

Strongly disagree
Somewhat disagree
Neither agree nor disagree
Somewhat agree
Strongly agree

Return