

Crafting Online Experiments

Thomas J. Leeper

Government Department
London School of Economics and Political Science

6 February 2018
KU-Leuven #MethLab

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.

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?

History/Logic

Theory

Challenges

Conclusion

Enter your data

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

History/Logic

Theory

Challenges

Conclusion

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

- Translating Hypotheses into Designs
- Assessing Quality
- Common Paradigms and Examples
- More Advanced Designs

3 Challenges and Criticisms

- Participant Recruitment
- Attention and Satisficing
- Use of Covariates

4 Conclusion

Who am I?

- Thomas Leeper
- Originally from Minnesota, USA
- Associate Professor in Political Behaviour at London School of Economics
- Research interests:
 - Survey experiments
 - Public opinion
 - 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 are your research interests?

Slides

Slides for the workshop are available at:

[http://thomasleeper.com/surveyexpcourse/
2018-leuven.html](http://thomasleeper.com/surveyexpcourse/2018-leuven.html)

1 History and Logic of Experiments

2 From Theory to Design

- Translating Hypotheses into Designs
- Assessing Quality
- Common Paradigms and Examples
- More Advanced Designs

3 Challenges and Criticisms

- Participant Recruitment
- Attention and Satisficing
- Use of Covariates

4 Conclusion

1 History and Logic of Experiments

2 From Theory to Design

- Translating Hypotheses into Designs
- Assessing Quality
- Common Paradigms and Examples
- More Advanced Designs

3 Challenges and Criticisms

- Participant Recruitment
- Attention and Satisficing
- Use of Covariates

4 Conclusion

Experiments: History I

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 II

- “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.)

Experiments: History II

- “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.)
- Multiple origins in the social sciences

Experiments: History II

- “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.)
- Multiple origins in the social sciences
 - First randomized experiment by Peirce and Jastrow (1884)
 - Gosnell (1924)
 - LaLonde (1986)
 - Gerber and Green (2000)

Experiments: History III

- Rise of surveys in the behavioral revolution
 - Survey research not heavily experimental because interviewing was mostly paper-based
 - “Split ballots” (e.g., Schuman & Presser; Bishop)

Experiments: History III

- Rise of surveys in the behavioral revolution
 - Survey research not heavily experimental because interviewing was mostly paper-based
 - “Split ballots” (e.g., Schuman & Presser; Bishop)
- 1983: Merrill Shanks and the Berkeley Survey Research Center develop CATI

Experiments: History III

- Rise of surveys in the behavioral revolution
 - Survey research not heavily experimental because interviewing was mostly paper-based
 - “Split ballots” (e.g., Schuman & Presser; Bishop)
- 1983: Merrill Shanks and the Berkeley Survey Research Center develop CATI
- Mid-1980s: Paul Sniderman & Tom Piazza performed the first *modern* survey experiment¹
 - Then: the “first multi-investigator”
 - Later: Skip Lupia and Diana Mutz created TESS

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

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
- Great resource for survey experimental materials, designs, and data
- Funded by the U.S. National Science Foundation
- Anyone anywhere in the world can apply
- See also: LISS, Bergen's Citizen Panel, Gothenburg's Citizen Panel

The First Survey Experiment

Hadley Cantril (1940) asks 3000 Americans either:

- Yes
- No

The First Survey Experiment

Hadley Cantril (1940) asks 3000 Americans either:

Do you think the U.S.
should do more than it is
now doing to help
England and France?

- Yes
- No

The First Survey Experiment

Hadley Cantril (1940) asks 3000 Americans either:

Do you think the U.S.
should do more than it is
now doing to help
England and France?

- Yes
- No

Do you think the U.S.
should do more than it is
now doing to help
England and France in
their fight against Hitler?

- Yes
- No

The First Survey Experiment

Hadley Cantril (1940) asks 3000 Americans either:

Do you think the U.S.
should do more than it is
now doing to help
England and France?

- Yes: 13%
- No

Do you think the U.S.
should do more than it is
now doing to help
England and France in
their fight against Hitler?

- Yes
- No

The First Survey Experiment

Hadley Cantril (1940) asks 3000 Americans either:

Do you think the U.S.
should do more than it is
now doing to help
England and France?

- Yes: 13%
- No

Do you think the U.S.
should do more than it is
now doing to help
England and France in
their fight against Hitler?

- Yes: 22%
- No

The First Survey Experiment

Hadley Cantril (1940) asks 3000 Americans either:

Do you think the U.S.
should do more than it is
now doing to help
England and France?

- Yes: 13%
- No

Do you think the U.S.
should do more than it is
now doing to help
England and France in
their fight against Hitler?

- Yes: 22%
- No

The “Hitler effect” was $22\% - 13\% = 9\%$

Definitions I

- A randomized experiment 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

Definitions I

- A randomized experiment 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

- If we manipulate the thing we want to know the effect of (X), and control (i.e., hold constant) everything we do not want to know the effect of (Z), the only thing that can affect the outcome (Y) is X .

Definitions II

Definitions II

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

Definitions II

- A survey experiment is just an experiment that occurs in a survey context
 - As opposed to in the field or in a laboratory
- Can be in any mode (face-to-face, CATI, IVR, CASI, etc.)

Definitions II

- A survey experiment is just an experiment that occurs in a survey context
 - As opposed to in the field or in a laboratory
- Can be in any mode (face-to-face, CATI, IVR, CASI, etc.)
- May or may not involve a representative population
 - Mutz (2011): “population-based survey experiments”

Definitions II

Definitions II

Unit: A physical object at a particular point in time

Definitions II

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

Synonyms: manipulation, intervention, factor, condition, cell

Definitions II

Outcome: The variable we are trying to explain

Definitions II

Potential outcomes: The outcome value 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 II

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

This is what we want to know!

Definitions II

Average causal effect: Difference in mean outcomes between treatment groups

This is almost what we want to know!

Example

Example

Unit: Americans in 1940

Example

Unit: Americans in 1940

Outcome: Support for military intervention

Example

Unit: Americans in 1940

Outcome: Support for military intervention

Treatment: Mentioning Hitler versus not

Example

Unit: Americans in 1940

Outcome: Support for military intervention

Treatment: Mentioning Hitler versus not

Potential outcomes:

- 1** Support in “Hitler” condition
- 2** Support in control condition

Example

Unit: Americans in 1940

Outcome: Support for military intervention

Treatment: Mentioning Hitler versus not

Potential outcomes:

- 1** Support in “Hitler” condition
- 2** Support in control condition

Causal effect: Difference in support between the two question wordings for each respondent

Example

Unit: Americans in 1940

Outcome: Support for military intervention

Treatment: Mentioning Hitler versus not

Potential outcomes:

- 1 Support in “Hitler” condition
- 2 Support in control condition

Causal effect: Difference in support between the two question wordings for each respondent

- Individual treatment effect not observable!

Example

Unit: Americans in 1940

Outcome: Support for military intervention

Treatment: Mentioning Hitler versus not

Potential outcomes:

- 1 Support in “Hitler” condition
- 2 Support in control condition

Causal effect: Difference in support between the two question wordings for each respondent

- Individual treatment effect not observable!
- Average effect (ATE) is the mean-difference

Questions?

Why are experiments useful?

Why are experiments useful?

Causal inference!

Addressing Confounding

In observational research...

Addressing Confounding

In observational research...

- 1 Correlate a “putative” cause (X) and an outcome (Y), where X temporally precedes Y

Addressing Confounding

In observational research...

- 1** Correlate a “putative” cause (X) and an outcome (Y), where X temporally precedes Y
- 2** Identify all possible confounds (Z)

Addressing Confounding

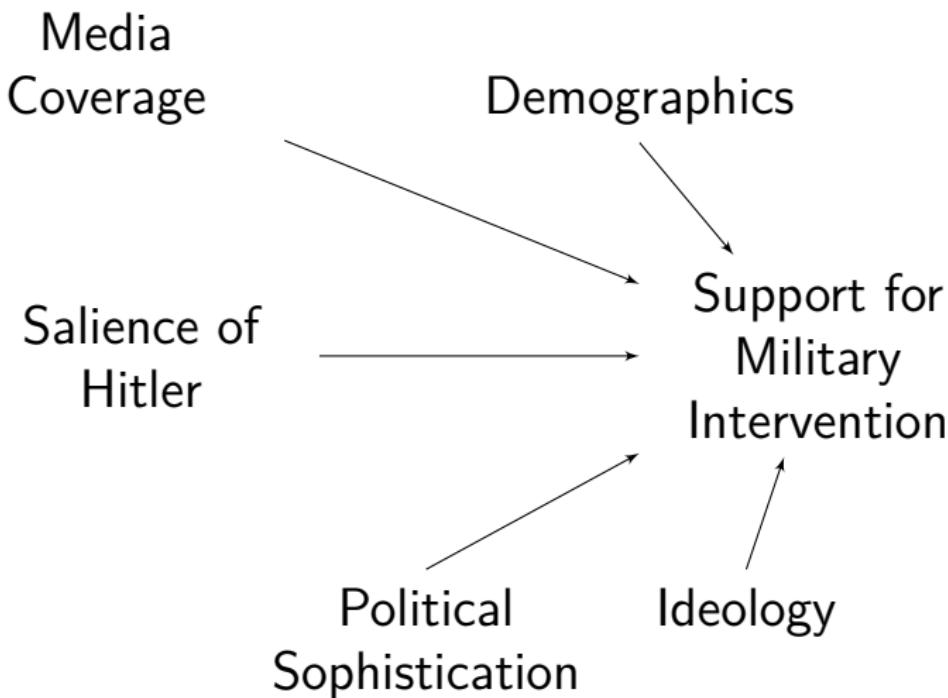
In observational research...

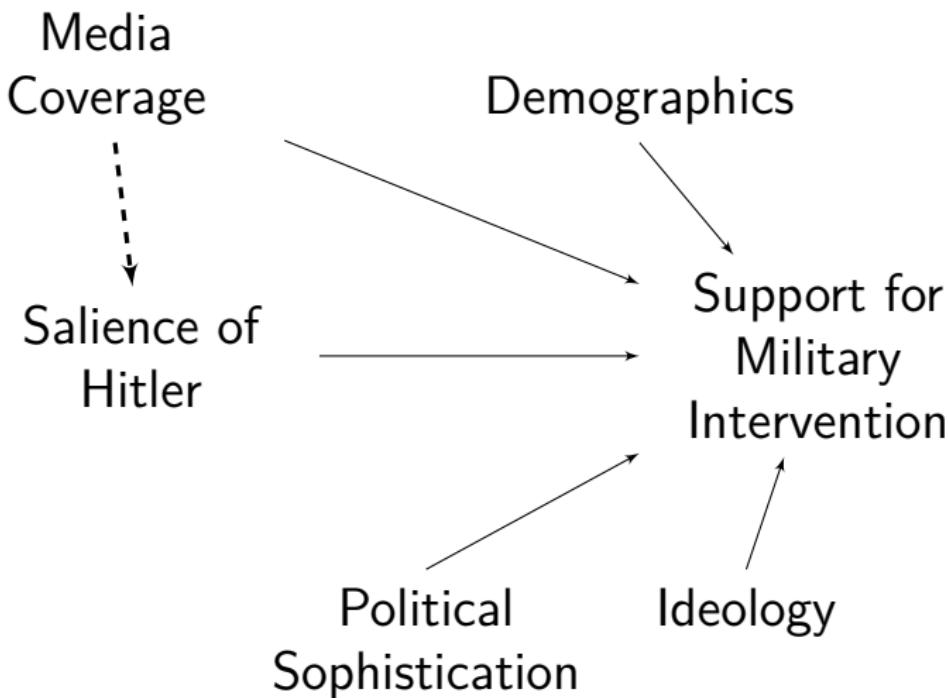
- 1 Correlate a “putative” cause (X) and an outcome (Y), where X temporally precedes Y
- 2 Identify all possible confounds (Z)
- 3 “Condition” on all confounds
 - Calculate correlation between X and Y at each combination of levels of Z

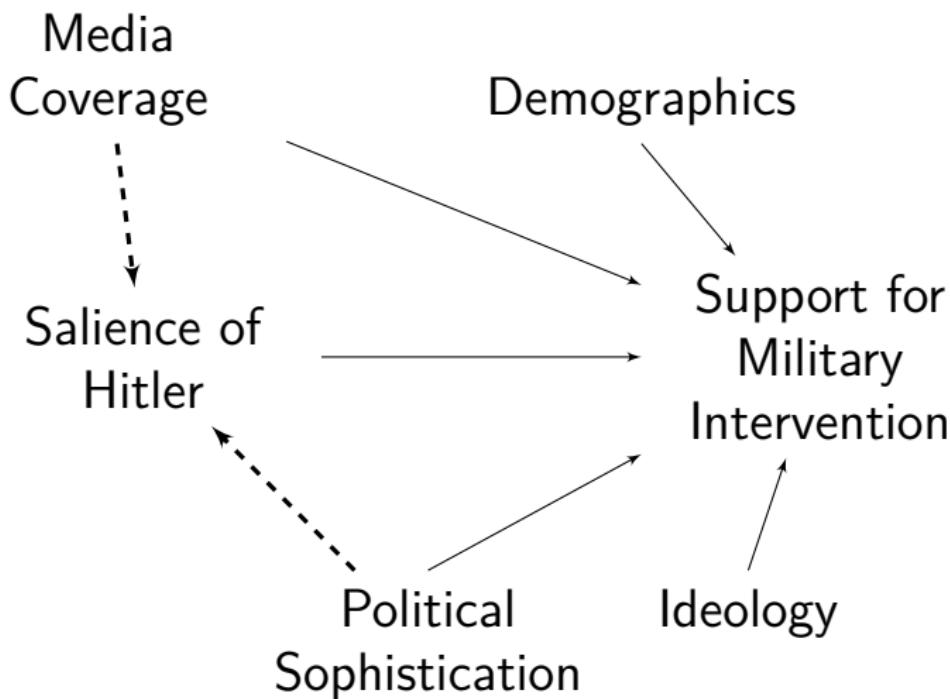
Addressing Confounding

In observational research...

- 1 Correlate a “putative” cause (X) and an outcome (Y), where X temporally precedes Y
- 2 Identify all possible confounds (Z)
- 3 “Condition” on all confounds
 - Calculate correlation between X and Y at each combination of levels of Z
- 4 Basically: $Y = \beta_0 + \beta_1 X + \beta_{2-k} Z + \epsilon$







Experiments are different

Experiments are different

- 1 Causal inferences from *design* not *analysis*

Experiments are different

- 1 Causal inferences from *design* not *analysis*
- 2 Solves both temporal ordering and confounding
 - Treatment (X) applied by researcher before outcome (Y)
 - Randomization eliminates confounding (Z)
 - We don't need to "control" for anything

Experiments are different

- 1 Causal inferences from *design* not *analysis*
- 2 Solves both temporal ordering and confounding
 - Treatment (X) applied by researcher before outcome (Y)
 - Randomization eliminates confounding (Z)
 - We don't need to "control" for anything
- 3 Basically: $Y = \beta_0 + \beta_1 X + \epsilon$

Experiments are different

- 1 Causal inferences from *design* not *analysis*
- 2 Solves both temporal ordering and confounding
 - Treatment (X) applied by researcher before outcome (Y)
 - Randomization eliminates confounding (Z)
 - We don't need to "control" for anything
- 3 Basically: $Y = \beta_0 + \beta_1 X + \epsilon$
- 4 Thus experiments are a "gold standard"

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.**

Questions?

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)

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 military support among those thinking of Hitler versus not

Experimental Inference II

- We cannot see individual-level causal effects
- We can see *average causal effects*
 - Ex.: Average difference in military support among those thinking of Hitler versus not
- 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:

$$ATE = E[Y_{1i}] - E[Y_{0i}] \quad (1)$$

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

Experimental Inference IV

- What we want and what we have:

$$ATE = E[Y_{1i}] - E[Y_{0i}] \quad (1)$$

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

- 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:

$$ATE = E[Y_{1i}] - E[Y_{0i}] \quad (1)$$

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

- 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:

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

Experimental Inference VI

- This holds in experiments because of a *physical process of randomization*²

²Random means “known probability of treatment” not “haphazard”.

Experimental Inference VI

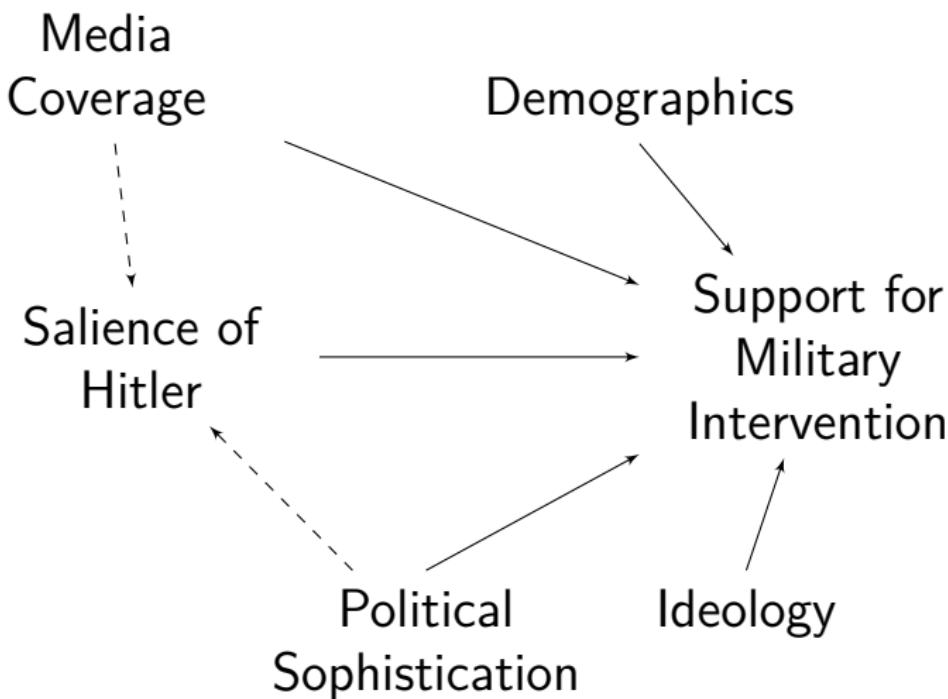
- This holds in experiments because of a *physical process of randomization*²
- Units differ only in side of coin that was up
 - $X_i = 1$ only because $D_i = 1$

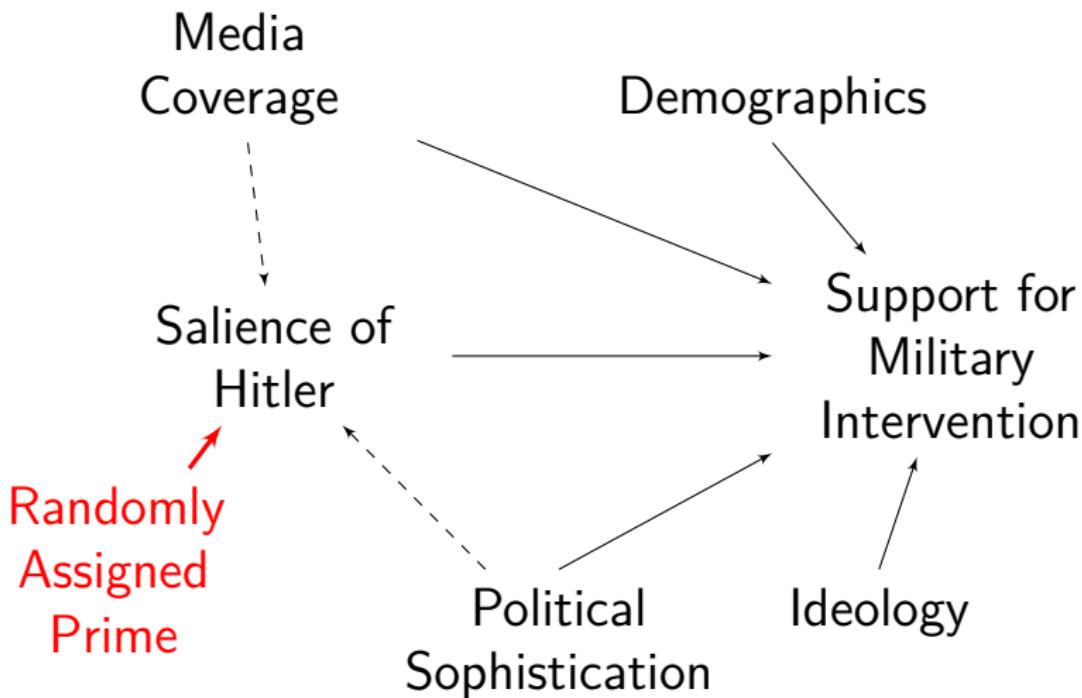
²Random means “known probability of treatment” not “haphazard”.

Experimental Inference VI

- This holds in experiments because of a *physical process of randomization*²
- 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)

²Random means “known probability of treatment” not “haphazard”.





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)
 - Design-based random sampling
 - Model-based re-weighting

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)
 - Design-based random sampling
 - Model-based re-weighting
- This boils down to being a mean-difference between two groups:

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

Tidy Experimental Data

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

Tidy Experimental Data

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

Tidy Experimental Data

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 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

R:

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

Stata:

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

Questions?

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)
- Thus we need to estimate the *variance* of the SATE
- The variance is influenced by:
 - Total sample size
 - Element variance of the outcome, Y
 - Relative size of each treatment group
 - (Some other factors)

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 → smaller SEs
- Smaller variance → 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 is used to determine sample size before conducting an experiment
- Type I and Type II Errors

	H_0 False ($ ATE > 0$)	H_0 True ($ATE = 0$)
Reject H_0	True positive	Type I Error
Accept H_0	Type II Error	True zero

- True positive rate ($1 - \kappa$) is power
- False positive rate is the significance threshold (α)

Doing a Power Analysis

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

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

Intuition about Power

Minimum detectable effect is the smallest effect we could detect given sample size, “true” ATE, variance of outcome measure, power ($1 - \kappa$), and α .

Intuition about Power

Minimum detectable effect is the smallest effect we could detect given sample size, “true” ATE, variance of outcome measure, power ($1 - \kappa$), and α .

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

Intuition about Power

Minimum detectable effect is the smallest effect we could detect given sample size, “true” ATE, variance of outcome measure, power ($1 - \kappa$), and α .

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

In underpowered study, we will be unlikely to detect true small effects. And most effects are small! ³

³Gelman, 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”
- Intuition: How large is the effect in standard deviations of the outcome?
 - Know if effects are large or small
 - Compare effects across studies

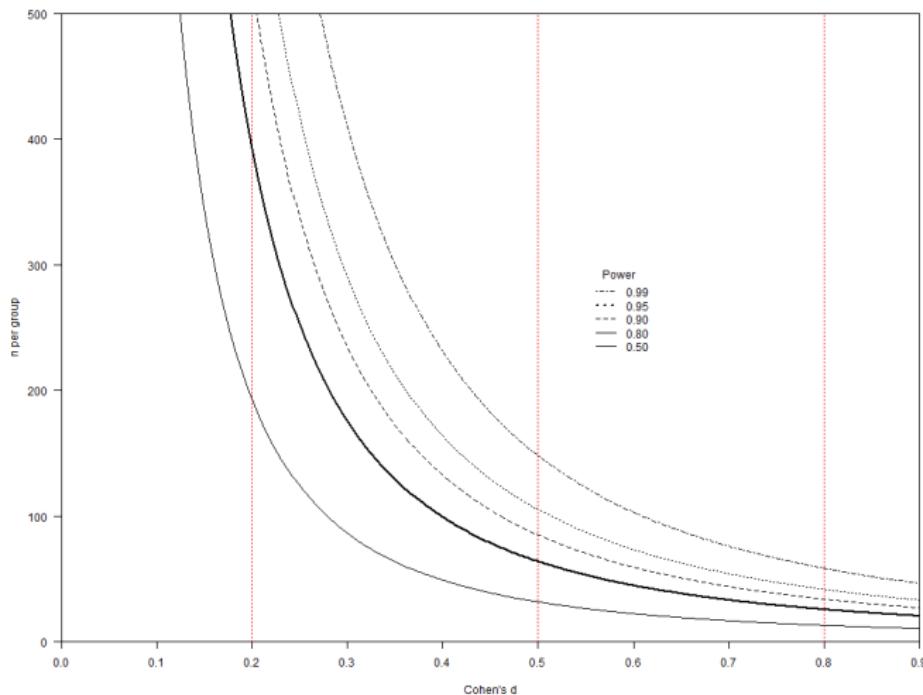
Intuition about Power

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

Intuition about Power

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

Intuition about Power



Power analysis in R

```
power.t.test(  
  # sample size (leave blank!)  
  n = ,  
  
  # minimum detectable effect size  
  delta = 0.4, sd = 1,  
  
  # alpha and power (1-kappa)  
  sig.level = 0.05, power = 0.8,  
  
  # two-tailed vs. one-tailed test  
  alternative = "two.sided"  
)
```

Power analysis in Stata

```
power twomeans 0, diff(0.2)
```

```
// for multiple values of  
forvalues i = 0.1 (0.1) 1.0 {  
    power twomeans 0, diff('i')  
}
```

```
// using raw effect sizes and standard deviations  
power twomeans 0 0.5, sd1(.5) sd2(.7)
```

```
// adjusting alpha or power  
power twomeans 0, diff(0.2) alpha(0.10) power(0.7)
```

Increasing/Decreasing Power

Increases Power

- Bigger sample
- Precise measures
- Covariates?

Decreases Power

- Attrition
- Noncompliance
- Clustering

History/Logic

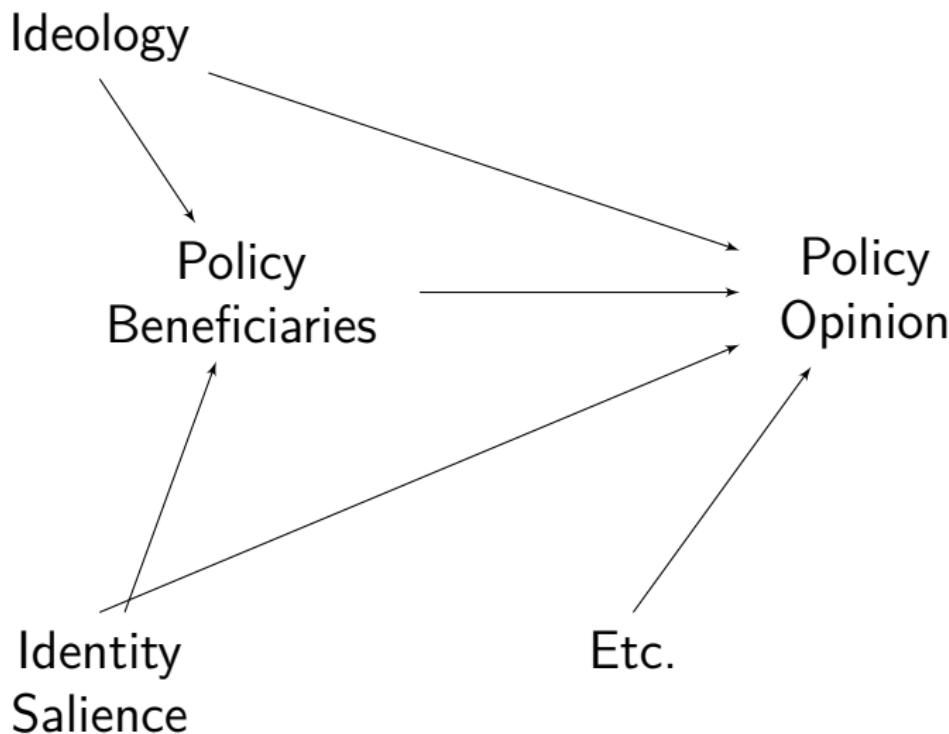
Theory

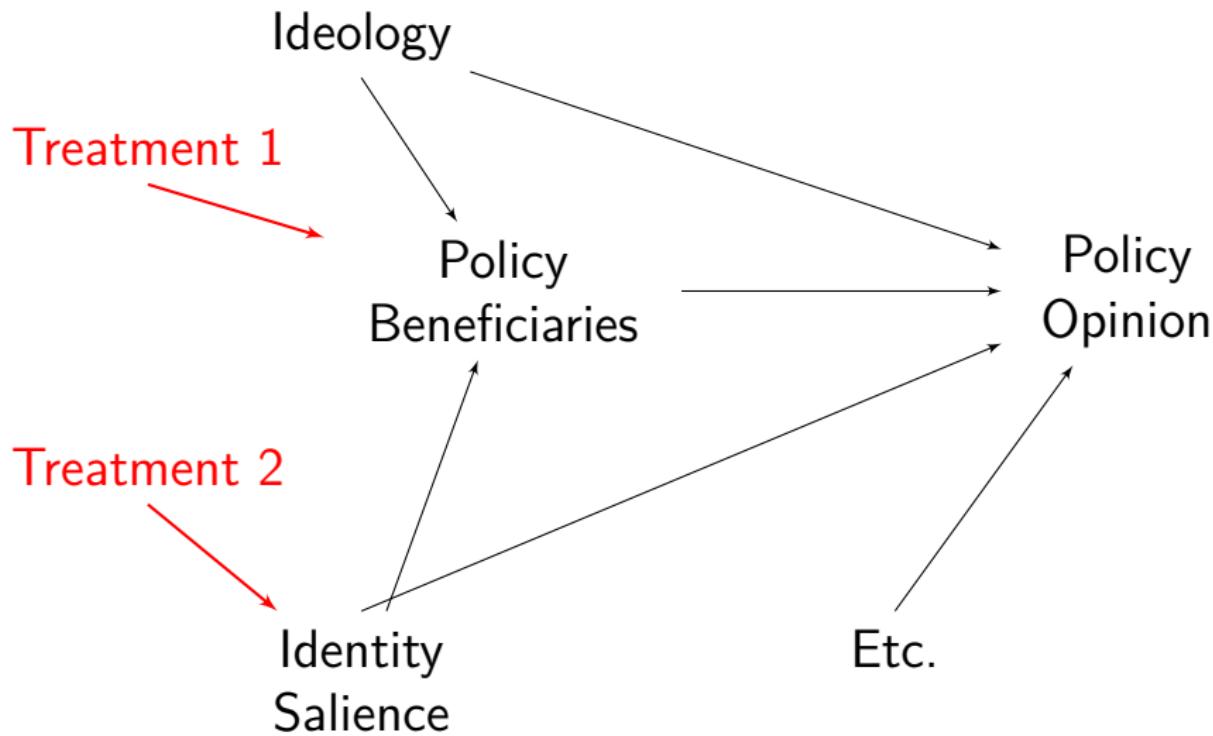
Challenges

Conclusion

Factorial Designs

- The two-condition experiment is a stylized ideal
- An experiment can have any number of conditions
 - Up to the limits of sample size
 - More than 8–10 conditions is typically unwieldy
- Three “flavors”:
 - Multiple conditions in a single factor
 - Multiple fully *crossed* factors
 - Partially crossed (“fractional factorial”) designs
- Regression methods provide a generalizable tool for causal inference in such designs





Example⁴

- 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?

⁴Transue. 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.

Example⁴

- 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?

⁴Transue. 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.

Example⁴

- 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?

⁴Transue. 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.

Example⁴

- 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?

⁴Transue. 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 *parameterize* this

Dummy variable regression (i.e., treatment–control CATEs):

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

Interaction effects (i.e., treatment–treatment CATEs):

$$Y = \beta_0 + \beta_1 X_{1,1} + \beta_2 X_{2,1} + \beta_3 X_{1,1} * X_{2,1} + \epsilon$$

Use margins to extract marginal effects

Considerations

- Factorial designs can quickly become unwieldy and expensive

Probably obvious, but . . .

Factors	Conditions per factor	Total Conditions	<i>n</i>
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

- Factorial designs can quickly become unwieldy and expensive

Considerations

- Factorial designs can quickly become unwieldy and expensive
- Need to consider what CATEs are of theoretical interest
 - Treatment-control, pairwise
 - Treatment-treatment, pairwise
 - Marginal effects, averaging across other factors
 - Comparison of merged conditions

Questions?

History/Logic

Theory

Challenges

Conclusion

1 History and Logic of Experiments

2 From Theory to Design

- Translating Hypotheses into Designs
- Assessing Quality
- Common Paradigms and Examples
- More Advanced Designs

3 Challenges and Criticisms

- Participant Recruitment
- Attention and Satisficing
- Use of Covariates

4 Conclusion

1 History and Logic of Experiments

2 From Theory to Design

- Translating Hypotheses into Designs
- Assessing Quality
- Common Paradigms and Examples
- More Advanced Designs

3 Challenges and Criticisms

- Participant Recruitment
- Attention and Satisficing
- Use of Covariates

4 Conclusion

From Theory to Design

- 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 ($H_0 = 0$)
- 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!

- Experimental “factors” are expressions of hypotheses as randomized groups

Treatments Test Hypotheses!

- Experimental “factors” are expressions of hypotheses as randomized groups
- What stimulus each group receives depends on hypotheses

Treatments Test Hypotheses!

- Experimental “factors” are expressions of hypotheses as randomized groups
- What stimulus each group receives depends on hypotheses
- Three ways hypotheses lead to stimuli:
 - 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?

1 History and Logic of Experiments

2 From Theory to Design

- Translating Hypotheses into Designs
- **Assessing Quality**
- Common Paradigms and Examples
- More Advanced Designs

3 Challenges and Criticisms

- Participant Recruitment
- Attention and Satisficing
- Use of Covariates

4 Conclusion

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 (5 February 2018)

**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 example⁵

- 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

⁵Leeper & 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)

⁶But, 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 front⁶

⁶But, consider the ethics of not doing so (more Friday)

1 History and Logic of Experiments

2 From Theory to Design

- Translating Hypotheses into Designs
- Assessing Quality
- **Common Paradigms and Examples**
- More Advanced Designs

3 Challenges and Criticisms

- Participant Recruitment
- Attention and Satisficing
- Use of Covariates

4 Conclusion

Question Wording Designs

- Simplest paradigm for presence/absence or qualitative variation
- Manipulation operationalizes this by asking two different questions
- Outcome is the answer to the question
- Example: Schuldt et al. “‘Global Warming’ or ‘Climate Change’? Whether the Planet is Warming Depends on Question Wording.”

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 example⁷

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)

⁷Singer & Couper. 2014. "The Effect of Question Wording on Attitudes toward Prenatal Testing and Abortion." *Public Opinion Quarterly* 78(3): 751–760.

Another framing example⁷

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)

⁷Singer & Couper. 2014. "The Effect of Question Wording on Attitudes toward Prenatal Testing and Abortion." *Public Opinion Quarterly* 78(3): 751–760.

Another framing example⁸

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

⁸Bobo & 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 example⁸

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?

⁸Bobo & 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 example⁹

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?

⁹ 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 example⁹

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?

⁹ 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 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-treatment¹⁰

- 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?

¹⁰Transue. 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-treatment¹⁰

- 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?

¹⁰Transue. 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-treatment¹⁰

- 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?

¹⁰Transue. 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-treatment¹⁰

- 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?

¹⁰Transue. 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?

An Instructional Manipulation¹¹

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.

¹¹Sturgis, Allum & Smith. 2008. "An Experiment on the Measurement of Political Knowledge in Surveys." *Public Opinion Quarterly* 72(1): 90–102.

An Instructional Manipulation¹¹

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.

¹¹Sturgis, Allum & Smith. 2008. "An Experiment on the Measurement of Political Knowledge in Surveys." *Public Opinion Quarterly* 72(1): 90–102.

An Instructional Manipulation + ¹²

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.

¹²Prior & 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 + ¹²

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.

¹²Prior & Lupia. 2008. "Money, Time, and Political Knowledge: Distinguishing Quick Recall and Political Learning Skills." *American Journal of Political Science* 52(1): 169–183.

Vignettes

- A “vignette” is a short 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 vignette¹³

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)

¹³Gilens, M. 1996. "'Race coding' and white opposition to welfare. *American Political Science Review* 90(3): 593–604.

Newer vignette¹⁴

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.

¹⁴Banerjee et al. 2012. "Are Poor Voters Indifferent to Whether Elected Leaders are Criminal or Corrupt? A Vignette Experiment in Rural India." Working paper.

Longer vignette example¹⁵

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

¹⁵Merolla & 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.

Longer vignette example¹⁵

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

¹⁵Merolla & 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

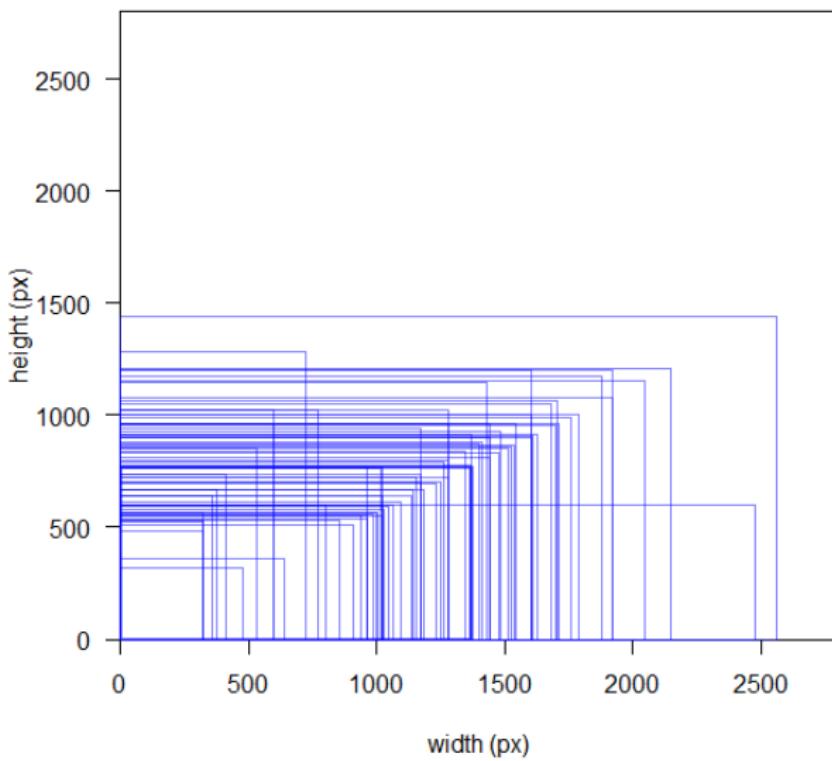
- Comparability across conditions
 - Length
 - Readability
- Language proficiency

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)



Non-textual Manipulations

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

¹⁶“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 & Gross¹⁶ measure impact of patriotic cues on candidate support by showing images of candidates with and without flags

¹⁶ "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 & Gross¹⁶ measure impact of patriotic cues on candidate support by showing images of candidates with and without flags
 - Subliminal primes possible, depending on software

¹⁶ "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 & Gross¹⁶ 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

¹⁶ "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.

Example¹⁷



Light Complexion



Original



Dark Complexion

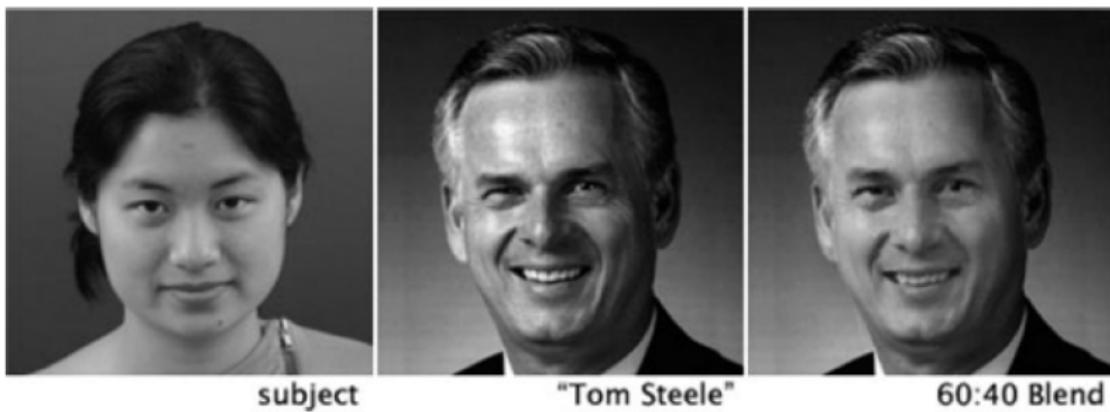
¹⁷Iyengar et al. 2010. "Do Explicit Racial Cues Influence Candidate Preference? The Case of Skin Complexion in the 2008 Campaign." Working paper.

Example¹⁸



¹⁸Laustsen & Petersen. 2016. "Winning Faces vary by Ideology." *Political Communication* 33(2): 188–211.

Example¹⁹



¹⁹ 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

²⁰Vavreck. 2007 "The Exaggerated Effects of Advertising on Turnout: The Dangers of Self-Reports." *Quarterly Journal of Political Science* 2: 325–343.

²¹Mutz. 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

²⁰Vavreck. 2007 "The Exaggerated Effects of Advertising on Turnout: The Dangers of Self-Reports." *Quarterly Journal of Political Science* 2: 325–343.

²¹Mutz. 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 Advertisements²⁰
 - News Programs²¹

²⁰Vavreck. 2007 "The Exaggerated Effects of Advertising on Turnout: The Dangers of Self-Reports." *Quarterly Journal of Political Science* 2: 325–343.

²¹Mutz. 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?

1 History and Logic of Experiments

2 From Theory to Design

- Translating Hypotheses into Designs
- Assessing Quality
- Common Paradigms and Examples
- More Advanced Designs

3 Challenges and Criticisms

- Participant Recruitment
- Attention and Satisficing
- Use of Covariates

4 Conclusion

Beyond Simple Designs

- 1 Factorial designs
- 2 Sensitive question designs
- 3 Conjoint designs
- 4 Multi-component designs
 - Over-time measurement/randomization
 - Field–survey combinations

Sensitive Item Designs

- Randomization can be used to measure something
- List experiments
 - Randomly present lists of items of varying length
 - Difference in count of items supported is prevalence of sensitive attitude/behavior
- Randomized response
 - Present a sensitive question
 - Use a randomization device to dictate whether the respondent answers the sensitive question or something else

List Experiments²²

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

²²Kuklinski et al. 1997. "Racial Prejudice and Attitudes Toward Affirmative Action." *American Journal of Political Science* 41(2): 402-419.

List Experiments²²

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

²²Kuklinski et al. 1997. "Racial Prejudice and Attitudes Toward Affirmative Action." *American Journal of Political Science* 41(2): 402-419.

Randomized Response²³

■ Example:

Here is a bag; in it there are stones from the game ‘Go,’ some colored black and others white. Please take one stone out, and see by yourself what color it is, black or white. Don’t let me know whether it is black or white, but be sure you know which it is. If you take a black one, answer the question: “Have you ever had an induced abortion?”

If you take a white one, answer the question: “Were you born in the lunar year of the horse?”

■ Considerations:

- Can use any randomization device
- Can be cognitively complex

²³Blair, Imai, and Zhou. 2015. “Design and Analysis of the Randomized Response Technique.” *JASA* 110(511): 1304–19.

Conjoint Analysis

- Surveys measure *stated* preferences
- Conjoint analysis involves measuring *revealed* preferences based upon a series of forced-choice decisions
 - Present respondents with pairs of “profiles” containing many *features*
 - Force respondents to choose which of the two they prefer
- Estimate *relative* importance of features of each profile

Advantages/Disadvantages

■ Advantages

- Reduces “cheap talk” results
- Lower social desirability biases
- Mimics real-world decisions
- Revealed preferences are causally interpretable

■ Disadvantages

- More cognitively complex for respondents than traditional polling
- No straightforward “% support” statistics

Structure of Conjoint

- Three examples:
 - 1 Policy preference on Brexit negotiations
 - 2 Choice of BBC Director General
 - 3 Choice of a lodger
- All are binary, forced-choice designs
- Analysis is all focused on AMCEs or subgroup AMCEs
 - Estimated using OLS dummy variable regression

Conjoint 1: Brexit Negotiations

YouGov

We are interested in your opinions about the negotiations between Britain and the European Union regarding Britain's exit from the EU and future relationship with the EU.

Please look carefully at these two possible outcomes:

	Outcome A	Outcome B
Britain's one-off payment to the EU to settle outstanding commitments	No payment	£10 billion
When this will come into effect	2025	2023
Border checks between Northern Ireland and the Republic of Ireland	No passport checks and no customs checks	Full passport and customs checks
EU's legal authority in Britain	Britain adopts some EU laws but is not subject to decisions by the European Court of Justice	Britain is subject to all EU laws and all decisions by the European Court of Justice
Britain's future payments to the EU budget to access science and regional development programmes	£1 billion per year for access	£1 billion per year for access
Trade agreement with the EU	Many administrative barriers to trade in goods and services and 5% average tariff on goods	Few administrative barriers to trade in goods and services and 2.5% average tariff on goods
Policy on immigration from the EU	Full control over EU immigration and little to no EU immigration	Some control over EU immigration and lower levels of EU immigration than now
Future rights of current EU nationals in Britain and British nationals in the EU	All can stay indefinitely	Must apply for 'leave to remain' under the same terms as people from non-EU countries

Which of these two outcomes do you prefer?

Outcome A

Conjoint 2: BBC Director

Imagine that you are deciding who to appoint as the next Director General of the BBC. You have received the following information about two applicants and need to make a decision between them.

- | | |
|--|--|
| <ul style="list-style-type: none">- Tom- 68 years old- Has worked 21 years for the BBC- Has a degree from the University of Oxford- Didn't vote at the 2017 election- Voted Remain in the EU referendum- Former lawyer | <ul style="list-style-type: none">- Claire- 35 years old- Has never worked for the BBC- Has a PhD from the University of Exeter- Voted Conservative at the 2017 election- Didn't vote in the EU referendum- Former television producer |
|--|--|

Which of the two applicants would you prefer as the next Director General of the BBC?

Conjoint 3: Lodger

Imagine that you have a spare room that you want to rent out to a lodger. You have received the following information about two possible lodgers and need to make a decision between them.

- James
- 19 years old
- Full-time student
- Helps out at the local Anglican church
- Didn't vote at the 2017 election
- Voted Remain in the EU referendum
- Likes watching rugby
- Becky
- 35 years old
- Works for a private company
- Volunteers at an Oxfam shop
- Voted Conservative at the 2017 election
- Didn't vote in the EU referendum
- Likes playing videogames

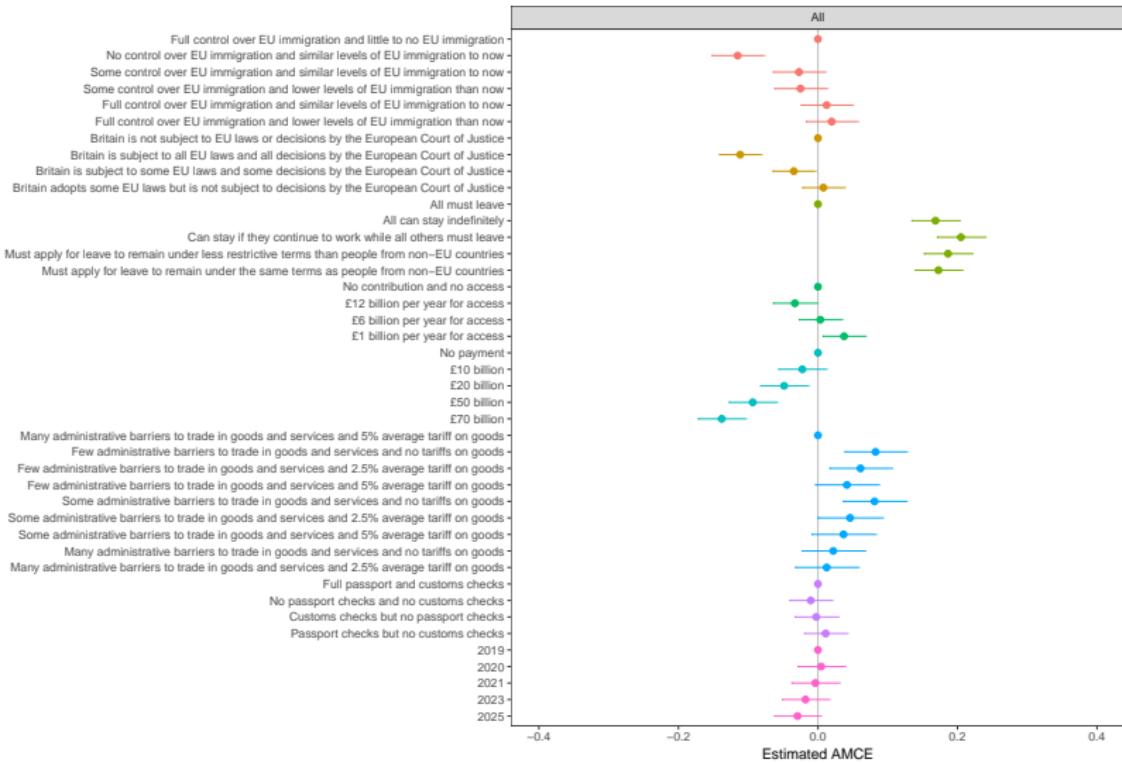
Which of the two lodgers would you prefer?

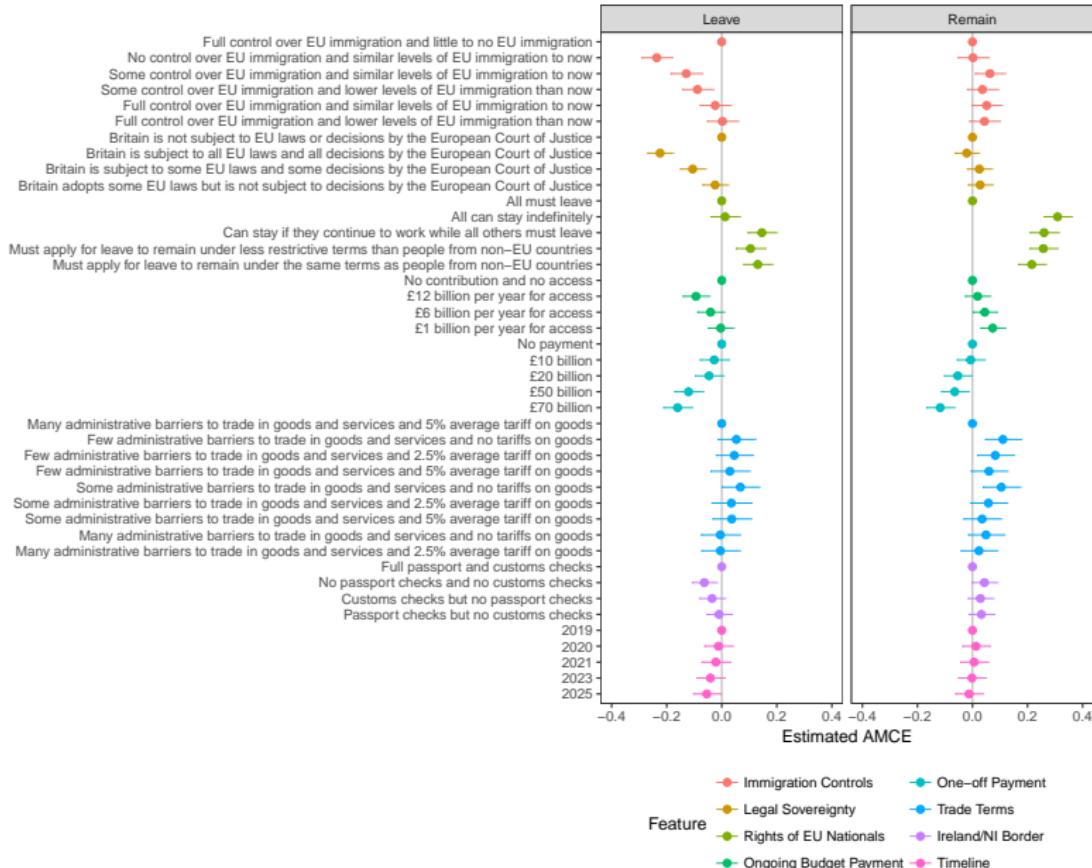
AMCEs

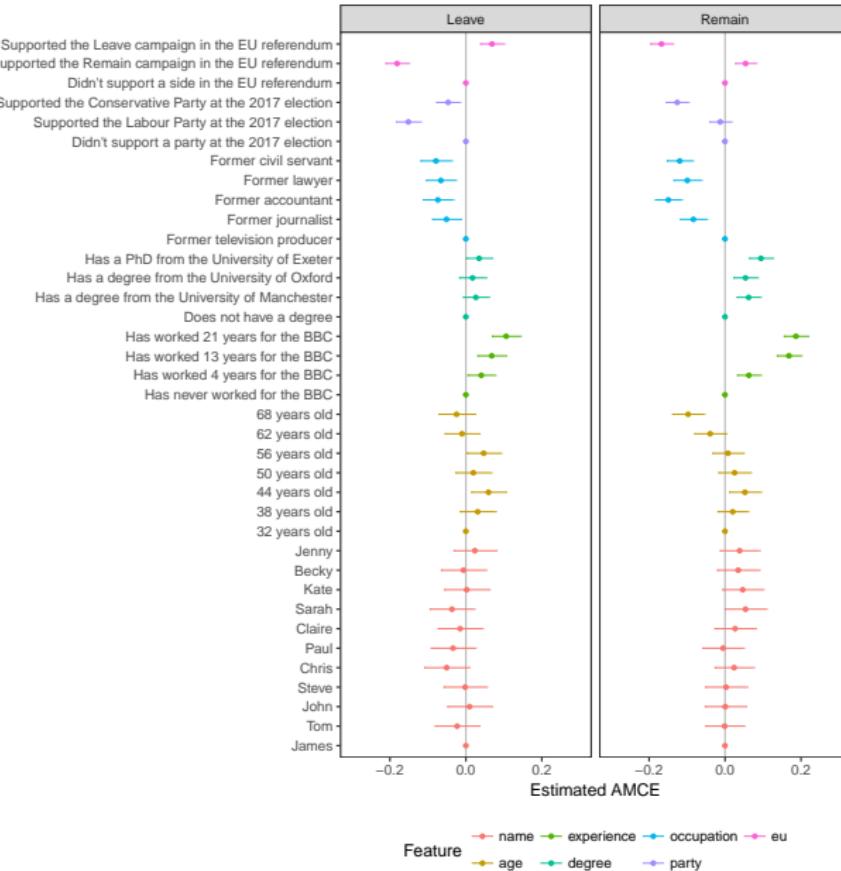
Statistic of interest is the *average marginal component effect* (AMCE), which is the causal effect of each level of each feature on support for an overall profile.

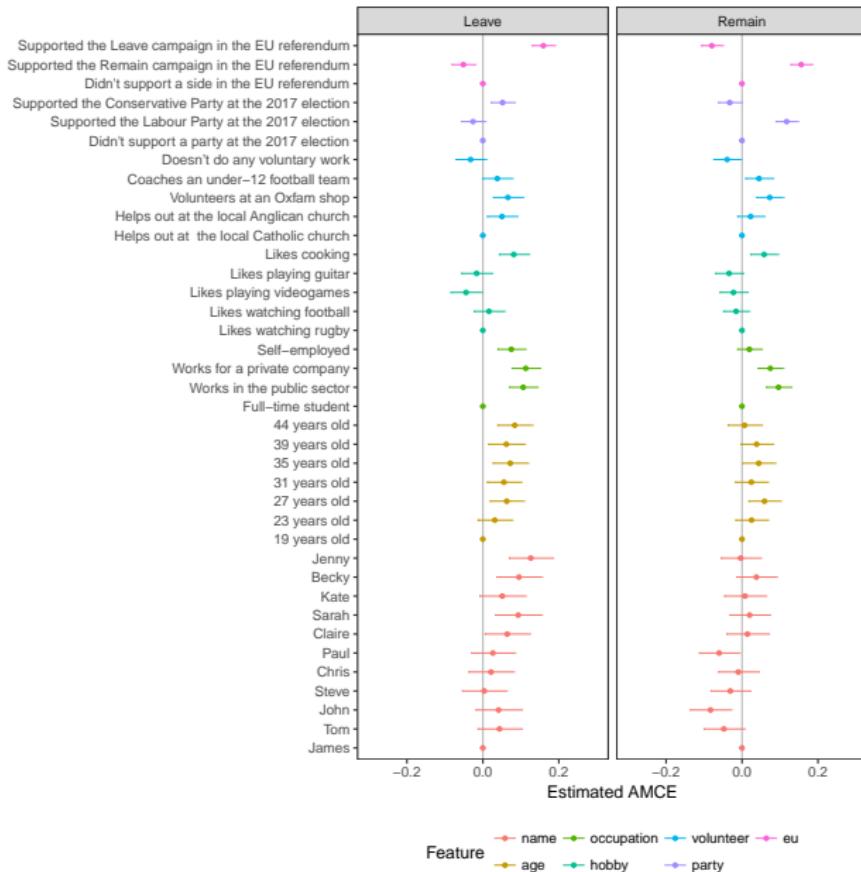
We can estimate this using (dummy variable) OLS, assuming:

- Full randomization of attributes and randomized pairing of profiles
- Even presentation of levels w/in features
- No profile ordering effects









Implementing a Conjoint

- Hope someone else can do it for you!
 - Requires programming
 - Not possible to manually create all possible combinations
- Strezhnev et al.'s tool:
<https://scholar.harvard.edu/astrezhnev/conjoint-survey-design-tool>
- Qualtrics using Javascript:
<https://github.com/leeper/conjoint-example>

Questions?

1 History and Logic of Experiments

2 From Theory to Design

- Translating Hypotheses into Designs
- Assessing Quality
- Common Paradigms and Examples
- More Advanced Designs

3 Challenges and Criticisms

- Participant Recruitment
- Attention and Satisficing
- Use of Covariates

4 Conclusion

1 History and Logic of Experiments

2 From Theory to Design

- Translating Hypotheses into Designs
- Assessing Quality
- Common Paradigms and Examples
- More Advanced Designs

3 Challenges and Criticisms

- Participant Recruitment
- Attention and Satisficing
- Use of Covariates

4 Conclusion

How do we find participants?

- Volunteers
 - Volunteer Science
 - In-house subject pool
- Paid crowdworkers
 - Prolific Academic
 - Mechanical Turk
 - Crowdflower
- “Representative” samples
 - Big players: YouGov, TNS, Gallup, Nielsen, GfK
 - Others: Kantar, SSI, Lucid

SUTO Framework

- Cronbach (1986) talks about generalizability in terms of UTO
- Shadish, Cook, and Campbell (2001) speak similarly of:
 - **Settings**
 - **Units**
 - **Treatments**
 - **Outcomes**
- 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

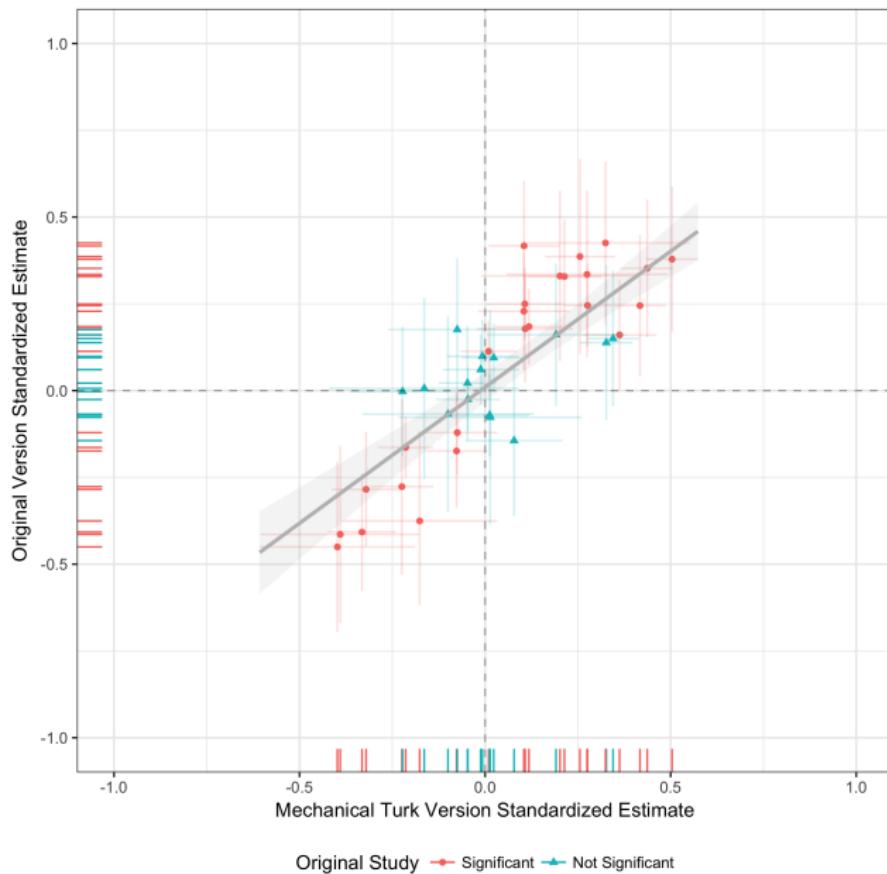
- Most common thing to focus on is demographic representativeness
 - Sears (1986): “students aren’t real people”
 - Western, educated, industrialized, rich, democratic (WEIRD) psychology participants

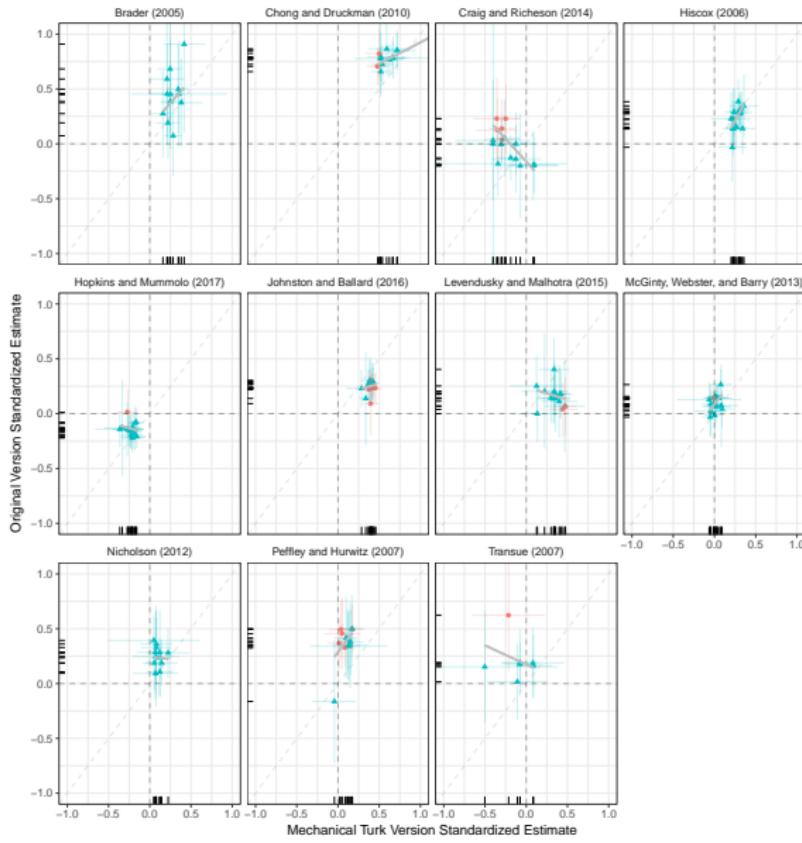
Common Differences

- 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





Questions?

1 History and Logic of Experiments

2 From Theory to Design

- Translating Hypotheses into Designs
- Assessing Quality
- Common Paradigms and Examples
- More Advanced Designs

3 Challenges and Criticisms

- Participant Recruitment
- Attention and Satisficing
- Use of Covariates

4 Conclusion

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: Pre-Analysis 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

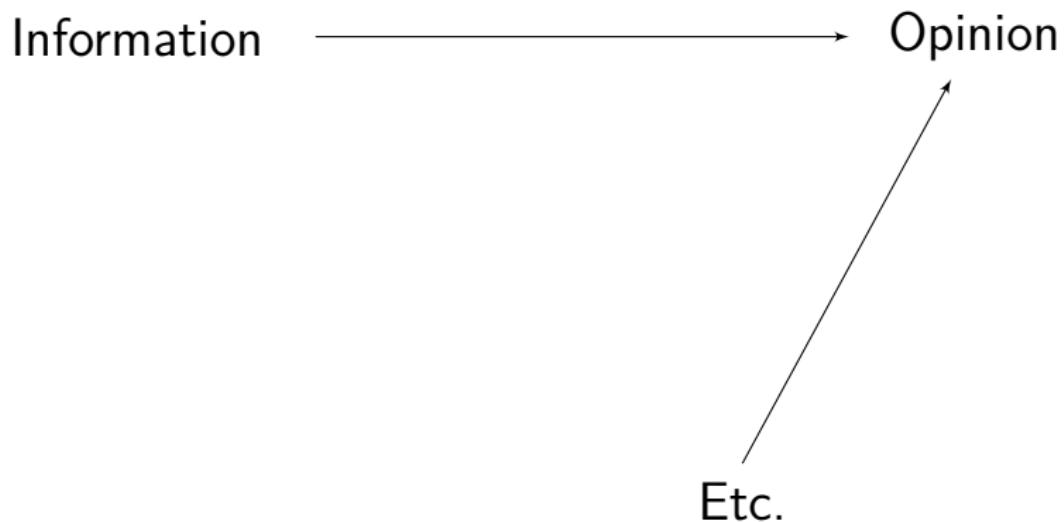
- This is totally fine from a causal inference perspective
- Advantages:
 - Focused on engaged respondents
 - Likely increase impact of treatment

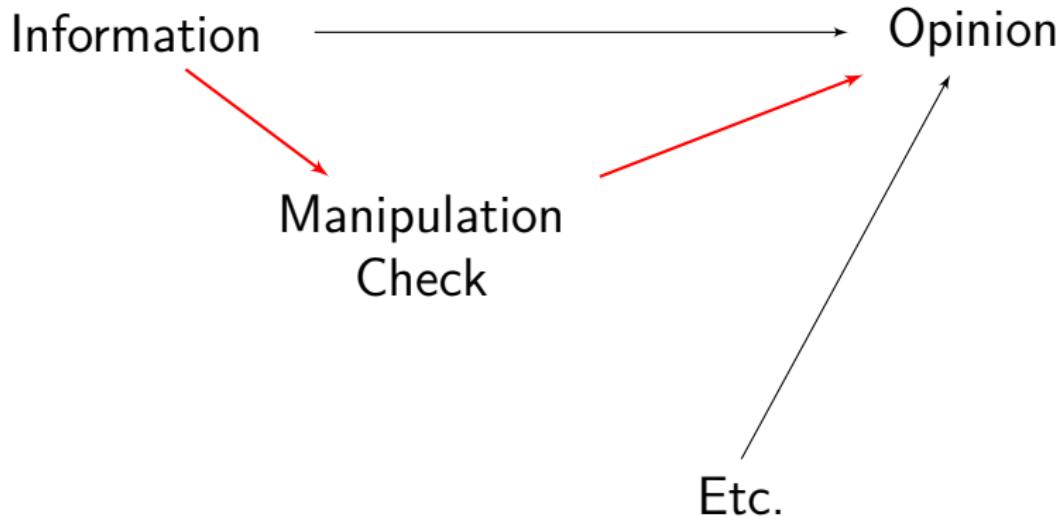
Pre-Treatment Exclusion

- 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)

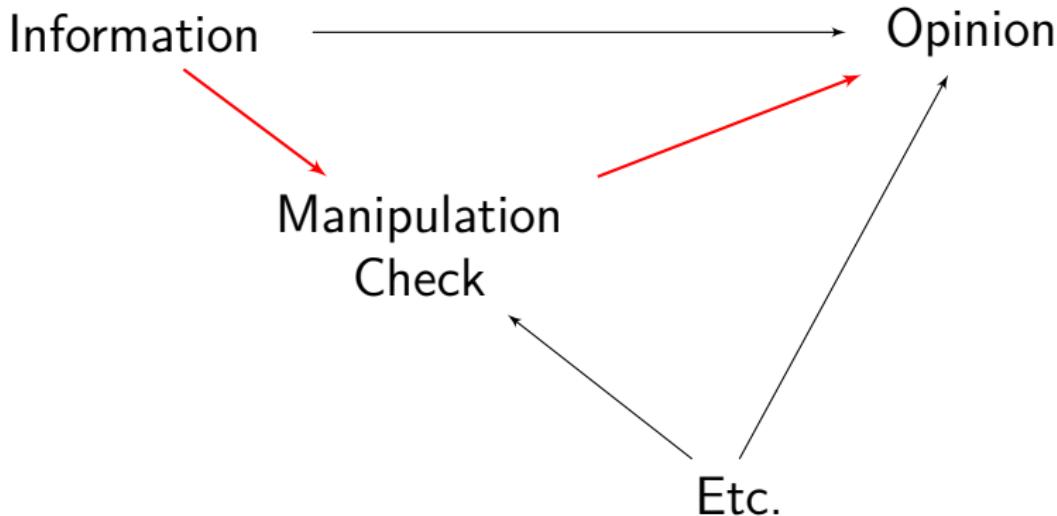
Post-Treatment Exclusion

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





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

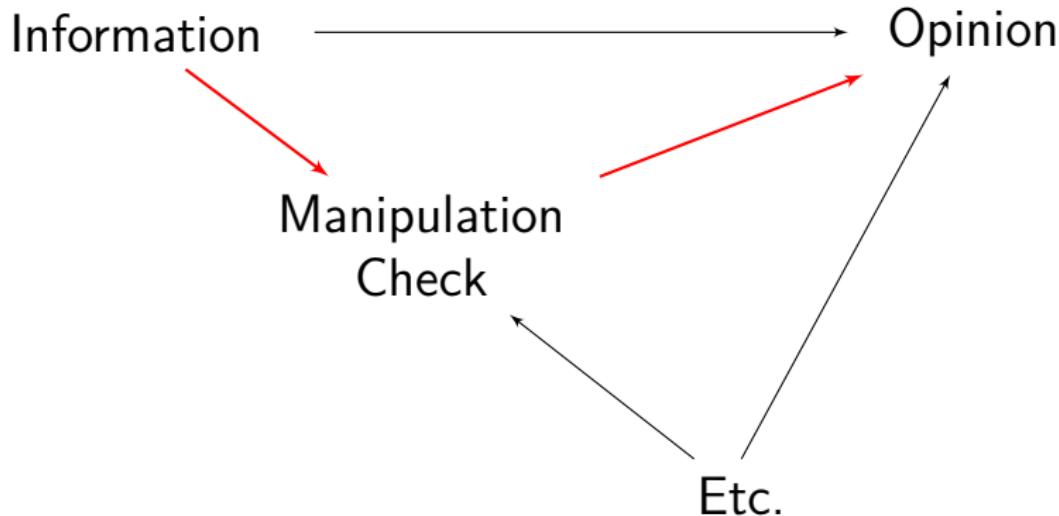
- Any post-treatment exclusion is problematic and should be avoided

Post-Treatment Exclusion

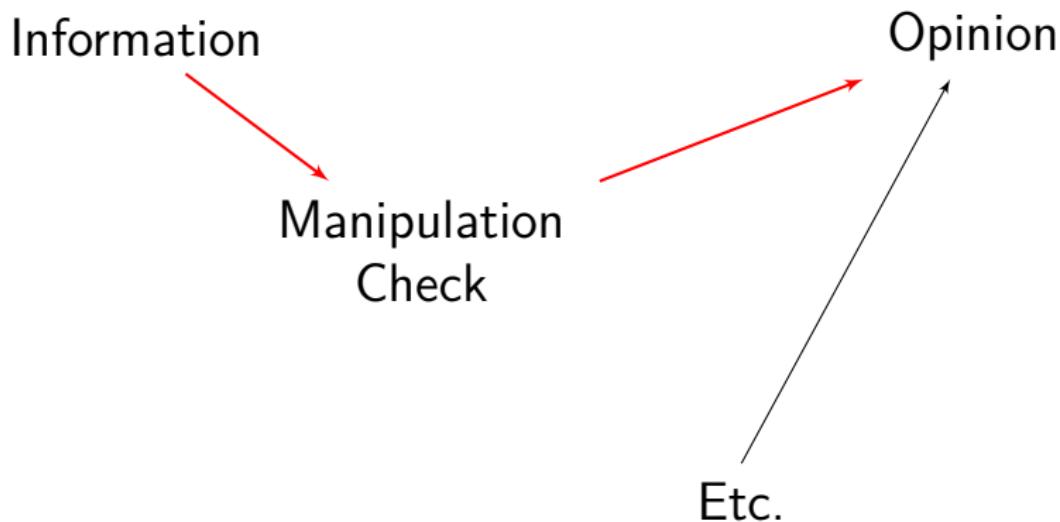
- Any post-treatment exclusion is problematic and should be avoided
- Can estimate a LATE
 - Interpretation: Effect of manipulation check among those whose value of the check can be changed by the treatment manipulation

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.



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?

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

Treatment Noncompliance

■ Definition:

“when subjects who were assigned to receive the treatment go untreated or when subjects assigned to the control group are treated”²⁴

²⁴Gerber & 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”²⁴

■ Several strategies

- “As treated” analysis
- “Intention to treat” analysis
- Estimate a LATE

²⁴Gerber & 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: $ITT = \bar{Y}_1 - \bar{Y}_0$

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: $ITT = \bar{Y}_1 - \bar{Y}_0$
- 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?

1 History and Logic of Experiments

2 From Theory to Design

- Translating Hypotheses into Designs
- Assessing Quality
- Common Paradigms and Examples
- More Advanced Designs

3 Challenges and Criticisms

- Participant Recruitment
- Attention and Satisficing
- Use of Covariates

4 Conclusion

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	
2	Male	Old	1	10	5
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	
2	Male	Old	1	10	5
3	Male	Young	1	4	
4	Male	Young	0	1	3
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
 - $\Pr(\text{Treated})$ 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?

1 History and Logic of Experiments

2 From Theory to Design

- Translating Hypotheses into Designs
- Assessing Quality
- Common Paradigms and Examples
- More Advanced Designs

3 Challenges and Criticisms

- Participant Recruitment
- Attention and Satisficing
- Use of Covariates

4 Conclusion

History/Logic

Theory

Challenges

Conclusion

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”?

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?

Types of Experiments

- 1 What are the three basic ways to construct experimental manipulations?

Types of Experiments

- 1 What are the three basic ways to construct experimental manipulations?
- 2 What are some useful and common paradigms for survey experiments?

Conjoints

1 What are conjoints useful for?

Conjoint

- 1 What are conjoints useful for?
- 2 How do we correctly analyze a conjoint experimental design?

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?

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!

History/Logic

Theory

Challenges

Conclusion

5 Protocols

6 Effect Heterogeneity

7 Advanced Designs

8 Beyond One-Shot Designs

TESS has “Open Protocols”

Protocol is the complete planning document for how to design, implement, and analyze an experiment.²⁵

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

²⁵Thomas 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

Detecting Effect Heterogeneity

Always block if you expect heterogeneity!

- QQ-plots: Suggestive evidence
- Regression using treatment-by-covariate interactions

Detecting Effect Heterogeneity

Always block if you expect heterogeneity!

- QQ-plots: Suggestive evidence
- Regression using treatment-by-covariate interactions
- (Replication and meta-analysis)

Suggestive Evidence

We can never know $\text{Var}(\text{TE}_i)$!

Suggestive Evidence

We can never know $\text{Var}(TE_i)$! But...

- Quantile-quantile plots

Suggestive Evidence

We can never know $\text{Var}(TE_i)$! 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 $\neq 1$

Suggestive Evidence

We can never know $\text{Var}(TE_i)$! But...

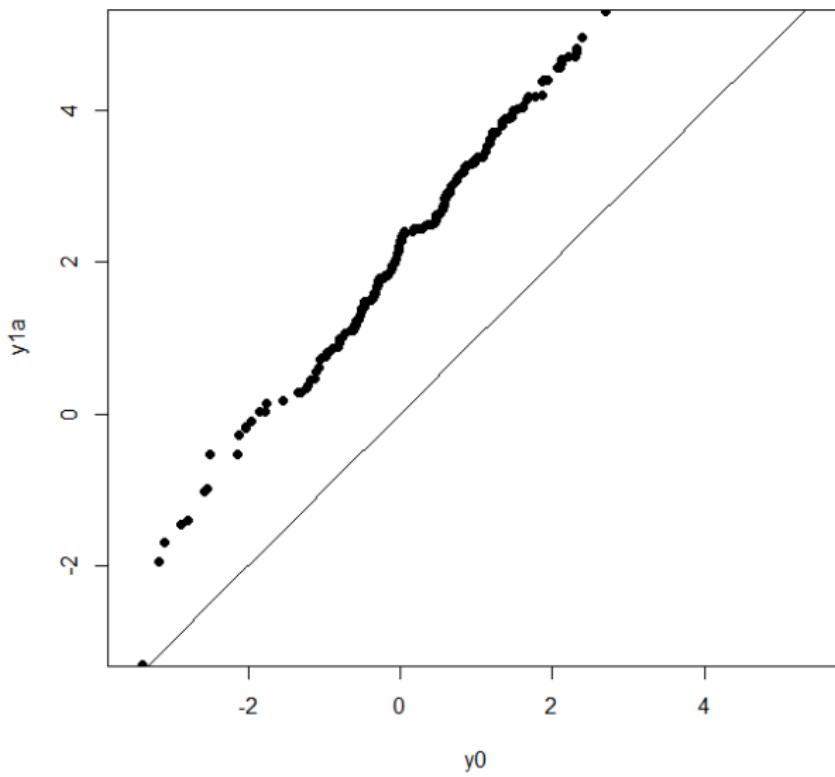
- Quantile-quantile plots
 - Compare the distribution of Y_0 's to distribution of Y_1 's
 - If homogeneity, a vertical shift in Y_1 's
 - If heterogeneity, a slope $\neq 1$
- Equality of variance tests
 - If homogeneity, variance should be equal
 - If heterogeneity, variances should differ

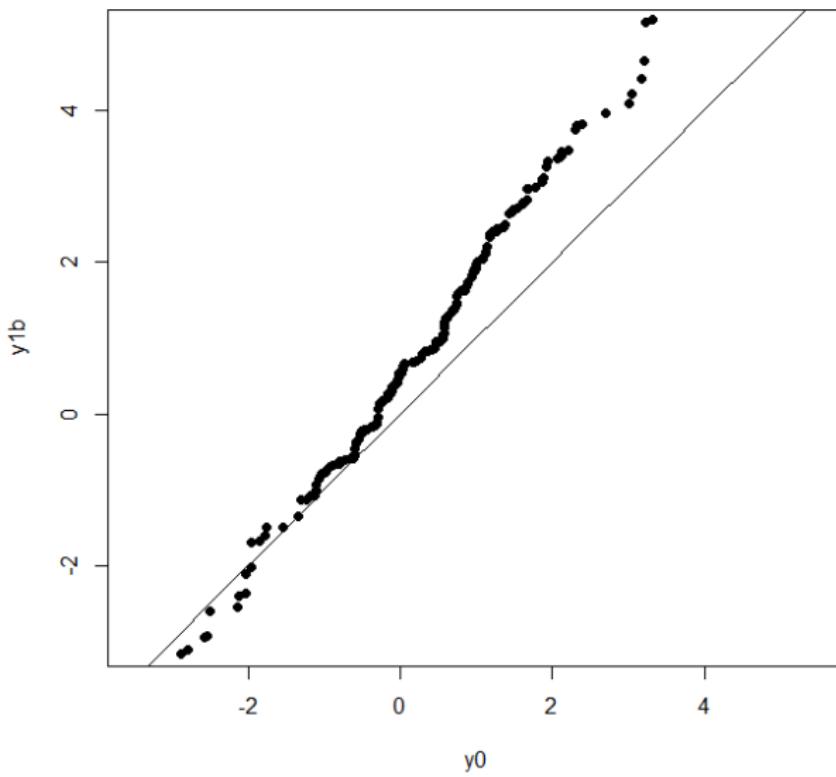
QQ Plots

```
# 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)
```





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: $\beta_3 \neq 0$

5 Protocols

6 Effect Heterogeneity

7 Advanced Designs

8 Beyond One-Shot Designs

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:

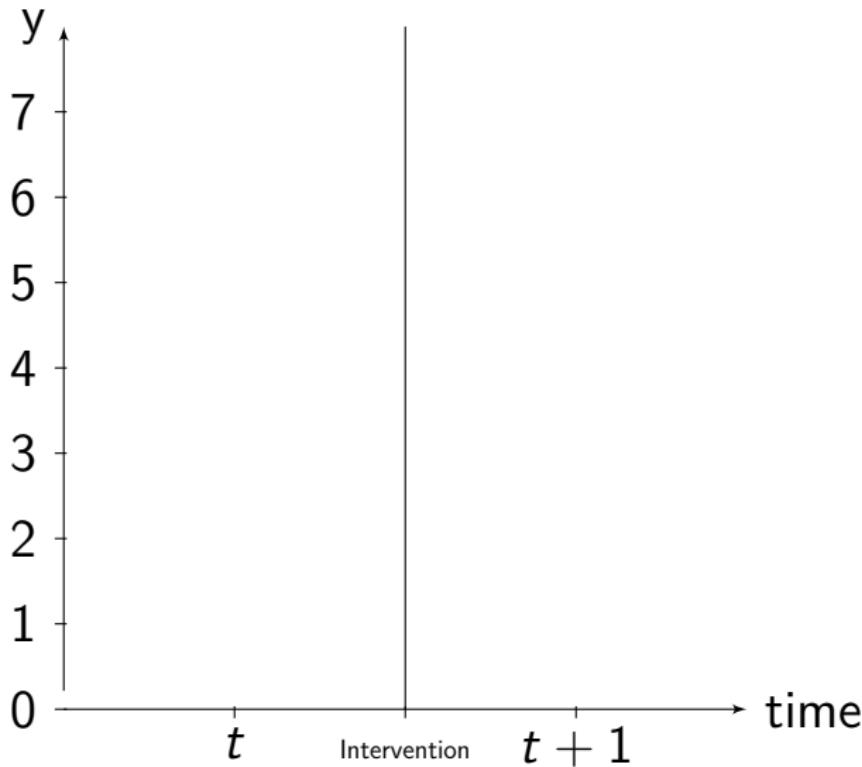
$$(\hat{Y}_{0,t+1} - \hat{Y}_{0,t}) - (\hat{Y}_{j,t+1} - \hat{Y}_{j,t})$$

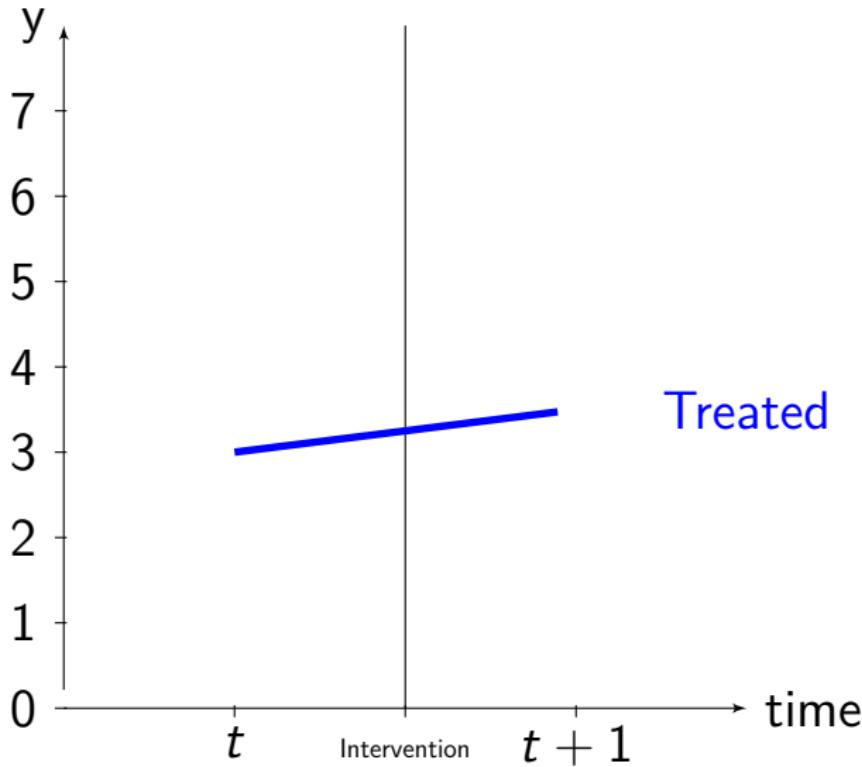
II. Within-Subjects Designs

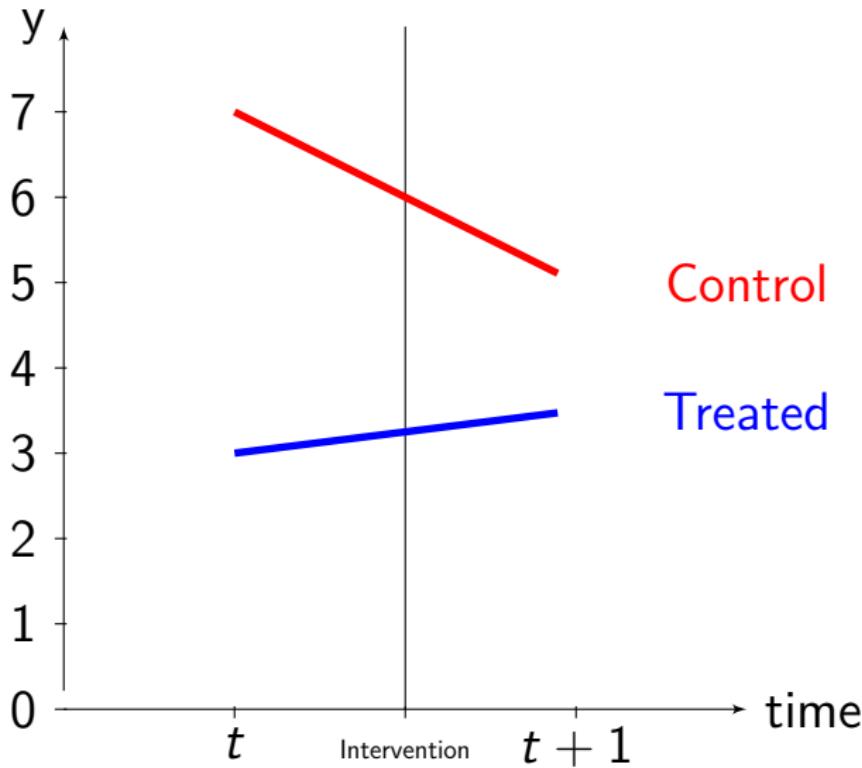
- Estimate treatment effects as a difference-in-differences
- Instead of using the post-treatment mean-difference in Y to estimate the causal effect, use the difference in pre-post differences for the two groups:

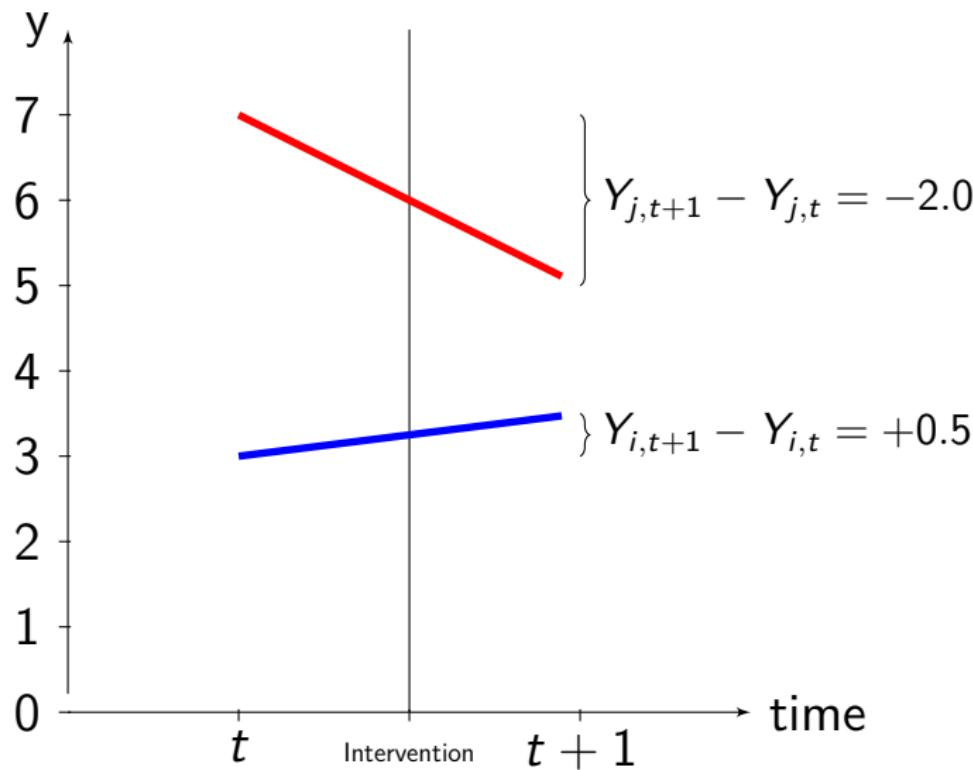
$$(\hat{Y}_{0,t+1} - \hat{Y}_{0,t}) - (\hat{Y}_{j,t+1} - \hat{Y}_{j,t})$$

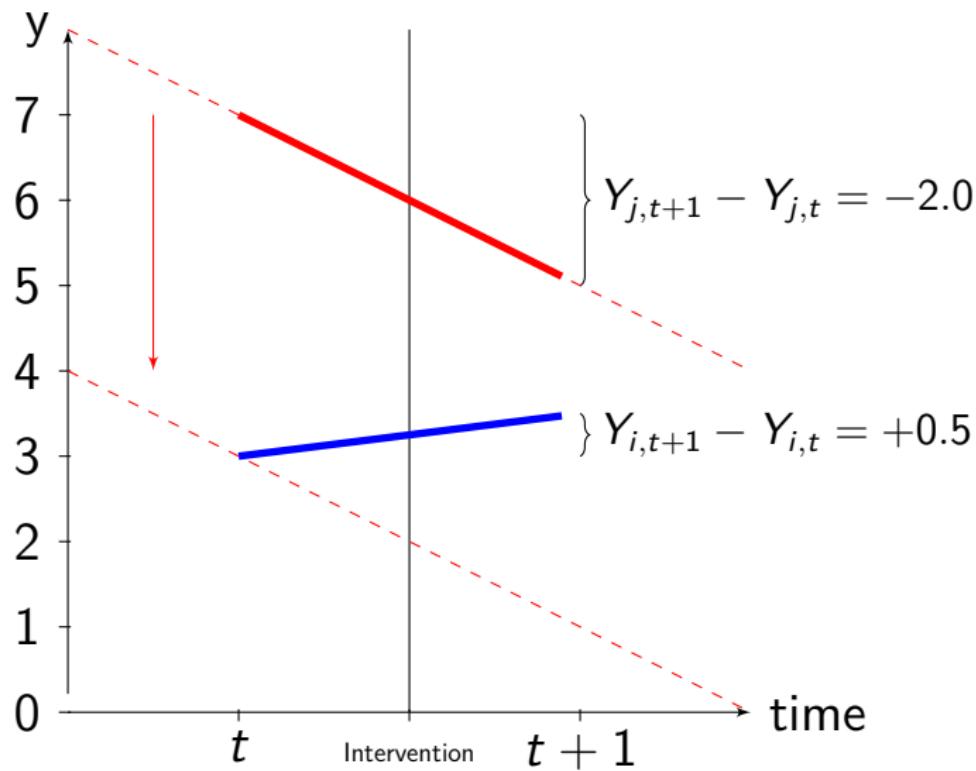
- Advantageous because variance for paired samples decreases as correlation between t_0 and t_1 observations increases

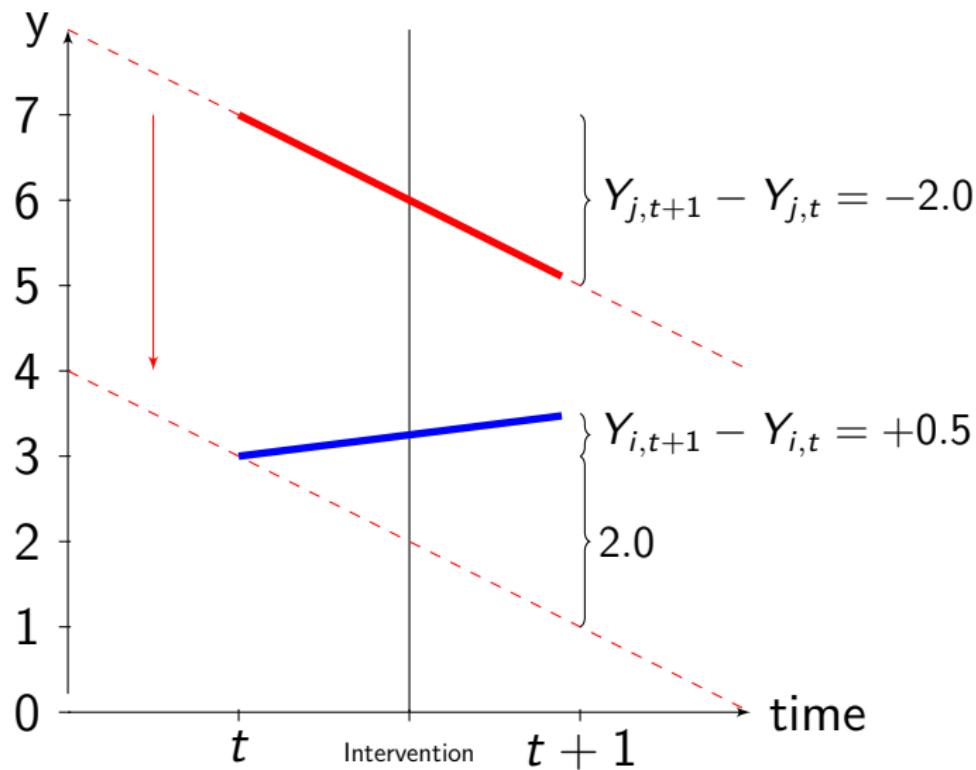


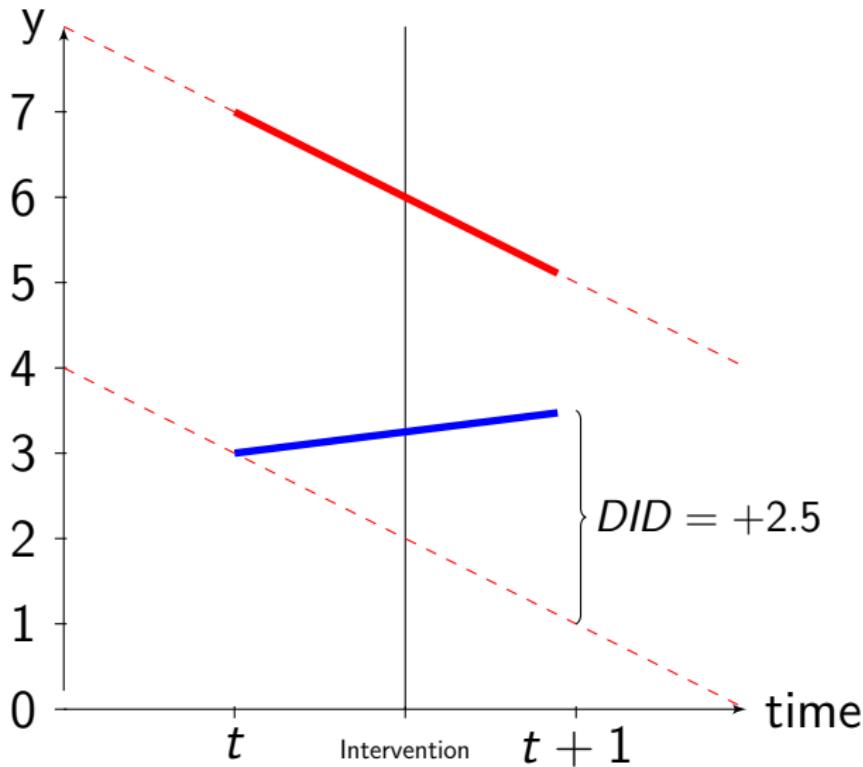












Threats to Validity

As soon as time comes into play, we have to worry about threats to validity.²⁶

²⁶Shadish, Cook, and Campbell (2002)

Threats to Validity

As soon as time comes into play, we have to worry about threats to validity.²⁶

- 1 History (simultaneous cause)

²⁶Shadish, Cook, and Campbell (2002)

Threats to Validity

As soon as time comes into play, we have to worry about threats to validity.²⁶

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

²⁶Shadish, Cook, and Campbell (2002)

Threats to Validity

As soon as time comes into play, we have to worry about threats to validity.²⁶

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

²⁶Shadish, Cook, and Campbell (2002)

Threats to Validity

As soon as time comes into play, we have to worry about threats to validity.²⁶

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

²⁶Shadish, Cook, and Campbell (2002)

Threats to Validity

As soon as time comes into play, we have to worry about threats to validity.²⁶

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

²⁶Shadish, Cook, and Campbell (2002)

Threats to Validity

As soon as time comes into play, we have to worry about threats to validity.²⁶

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

²⁶Shadish, 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 Lawrence²⁷

²⁷ 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 Lawrence²⁷
- Issues

²⁷ 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 Lawrence²⁷
- Issues
 - Nonresponse
 - Noncompliance

²⁷ Albertson & Lawrence. 2009. "After the Credits Roll." *American Politics Research* 37(2): 275–300.
10.1177/1532673X08328600.