Session I
Survey Experiments in Context

Thomas J. Leeper

Government Department
London School of Economics and Political Science

Who am I?

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

Who are you?

- Introduce yourself to a neighbour
- Where are you from?
- What do you hope to learn from the course?

Quick Survey

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

Course Materials

All material for the course is available at:

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

Learning Outcomes

By the end of the week, 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.

Schedule of Four Sessions

1. Survey Experiments in Context
2. Examples and Paradigms
3. Hands-on Session
4. Practical Issues

Questions?

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.)
- 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)
- 1983: Merrill Shanks and the Berkeley Survey Research Center develop CATI
- Mid-1980s: Paul Sniderman & Tom Piazza performed the first *modern* survey experiment[1]
  - Then: the "first multi-investigator"
  - Later: Skip Lupia and Diana Mutz created TESS

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

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:

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

- 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

- 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

Unit: A physical object at a particular point in time
Treatment: An intervention, whose effect(s) we wish to assess relative to some other (non-)intervention

Synonyms: manipulation, intervention, factor, condition, cell
Outcome: The variable we are trying to explain
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

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?

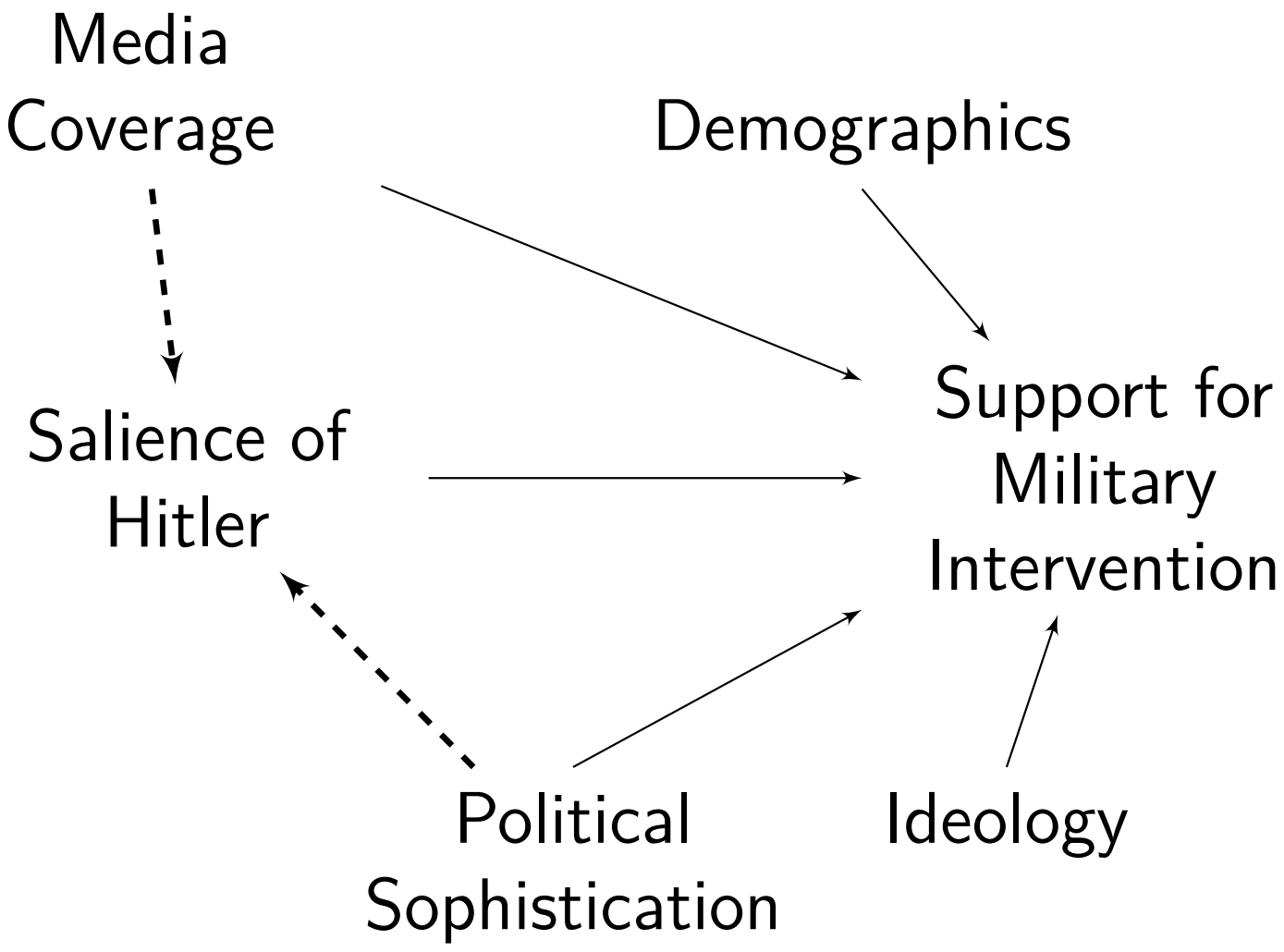
Why are experiments useful?

Causal inference!

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 ($\mathbf{Z}$)
3. "Condition" on all confounds
   - Calculate correlation between X and Y at each combination of levels of $\mathbf{Z}$
4. Basically: $Y = \beta_0 + \beta_1 X + \beta_{2-k}\mathbf{Z} + \epsilon$


Media Coverage
Demographics
Salience of Hitler
Support for Military Intervention
Political Sophistication
Ideology

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 ($\mathbf{Z}$)
    - 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**.

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

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

Experimental Inference IV

- What we want and what we have:

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

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

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

Experimental Inference V

- Only true when both of the following hold:

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

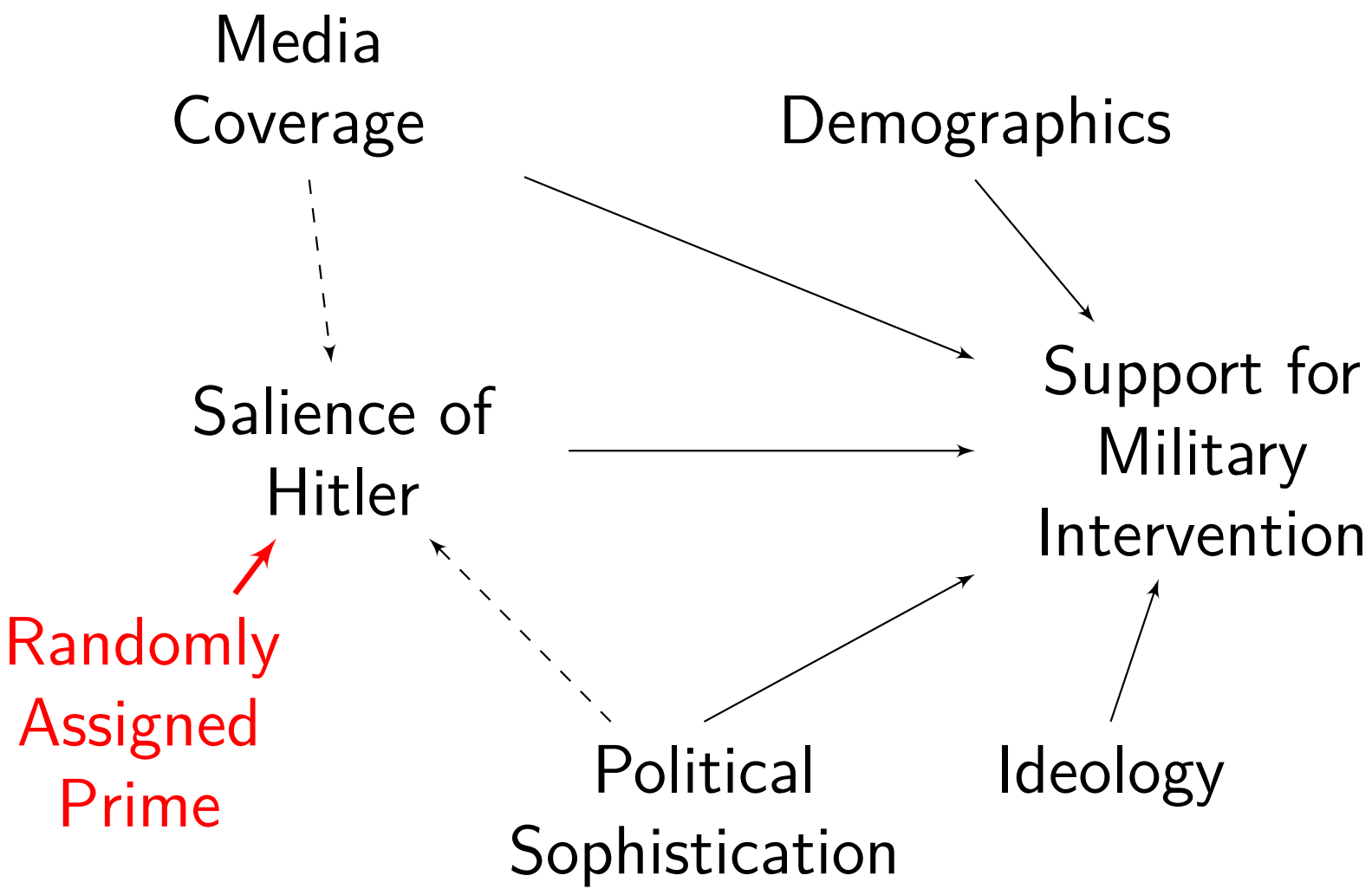
- In that case, potential outcomes are *independent* of treatment assignment
- If true (e.g., due to randomization of X), then:

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

Experimental Inference VI

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

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


Media Coverage
Demographics
Salience of Hitler
Support for Military Intervention
Randomly Assigned Prime
Political Sophistication
Ideology

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
- This boils down to being a mean-difference between two groups:

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

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

Statistical Power

- Power analysis is used to determine sample size before conducting an experiment
- Type I and Type II Errors

	H_0 False ($\lvert ATE \rvert > 0$)	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

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

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.

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

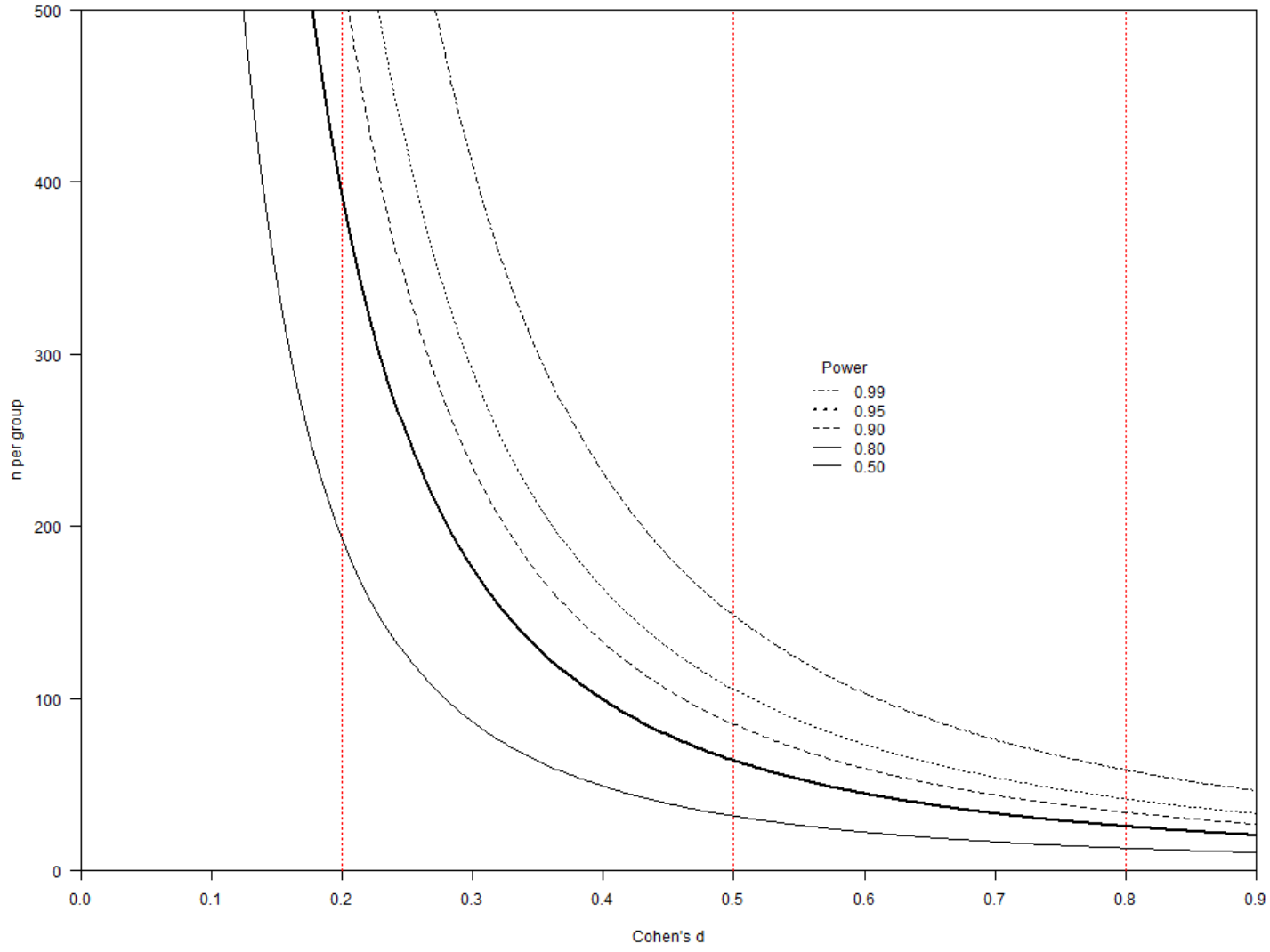
In underpowered study, we will be unlikely to detect true small effects. And most effects are small! [3]

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

Intuition about Power

- It can help to think in terms of "standardized effect sizes"
- 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:
  $d = \frac{\bar{x}_1 - \bar{x}_0}{s}$, 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

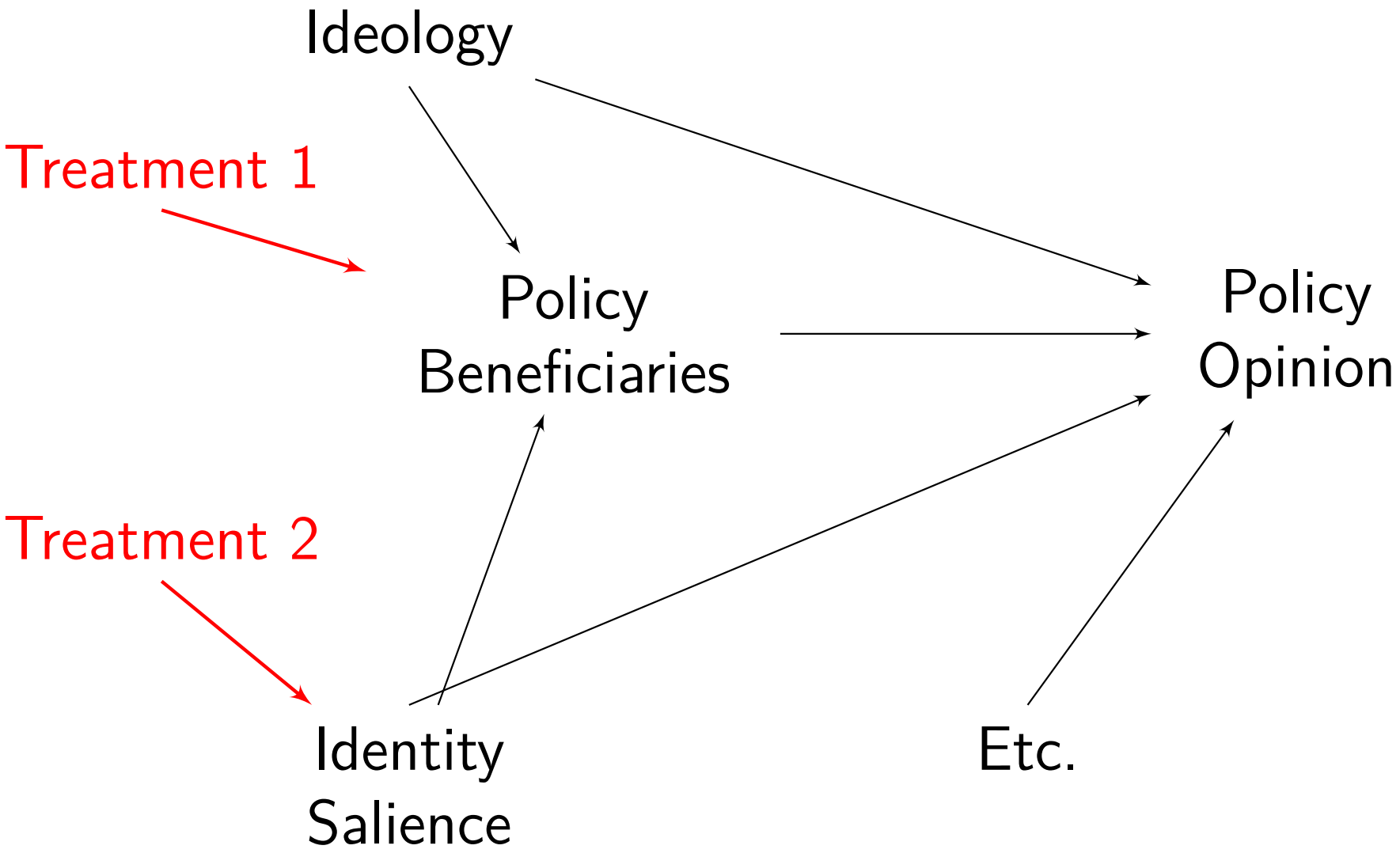
- Bigger sample
- Precise measures
- Covariates?

Decreases Power

- Attrition
- Noncompliance
- Clustering

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


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

Example[4]

- How close do you feel to your ethnic or racial group?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?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?

2x2 Factorial Design

Condition	
Educ. for Minorities	Y_1
Schools	Y_0

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):

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

Use `margins` to extract marginal effects

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

Probably obvious, but. . .

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

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

Considerations

- 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?