Session II
Survey Experiments in Practice

Thomas J. Leeper

Government Department
London School of Economics and Political Science

31 January 2017

Form groups of 3 and share the survey-experiment examples you found on TESS.

Beyond One-shot Designs

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

Beyond One-shot Designs

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

I. Effect Duration

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

II. Within-Subjects Designs

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

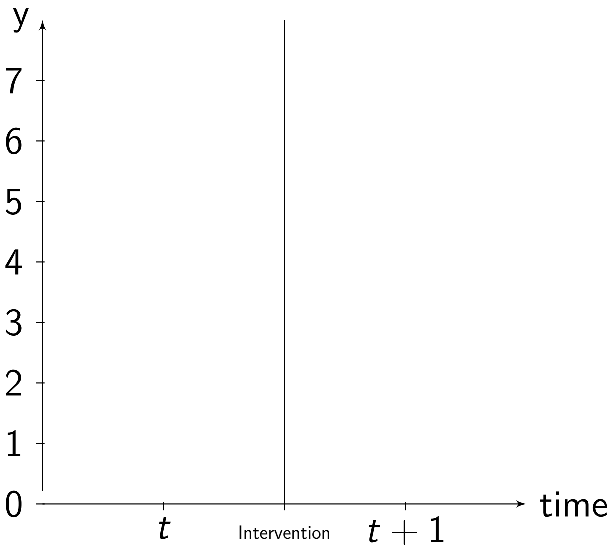
$$(\hat{Y}_{0,t+1} - \hat{Y}_{0,t}) - (\hat{Y}_{j,t+1} - \hat{Y}_{j,t})$$

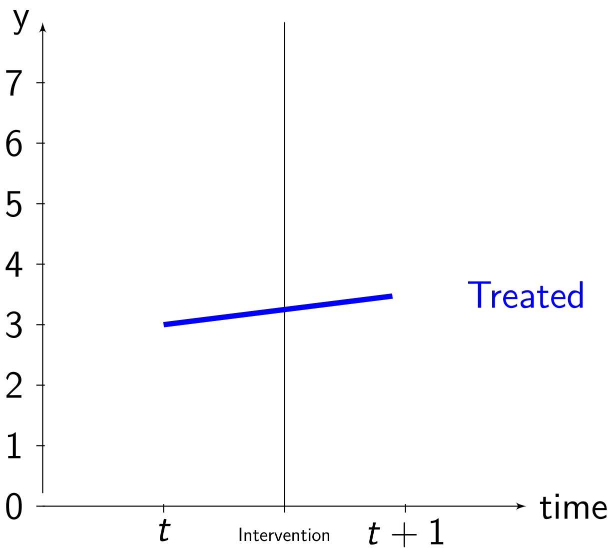
II. Within-Subjects Designs

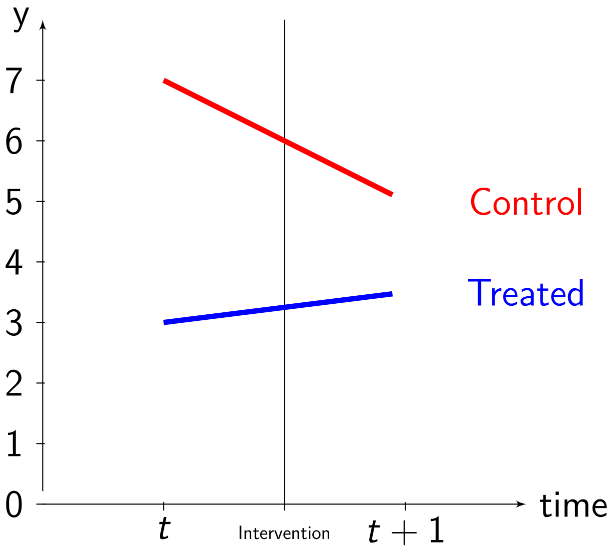
- Estimate treatment effects as a difference-in-differences
- Instead of using the post-treatment mean-difference in Y to estimate the causal effect, use the difference in pre-post differences for the two groups:

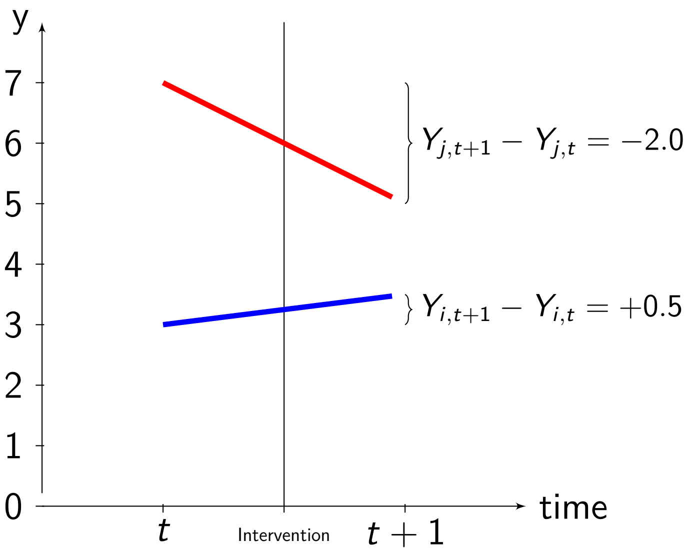
$$(\hat{Y}_{0,t+1} - \hat{Y}_{0,t}) - (\hat{Y}_{j,t+1} - \hat{Y}_{j,t})$$

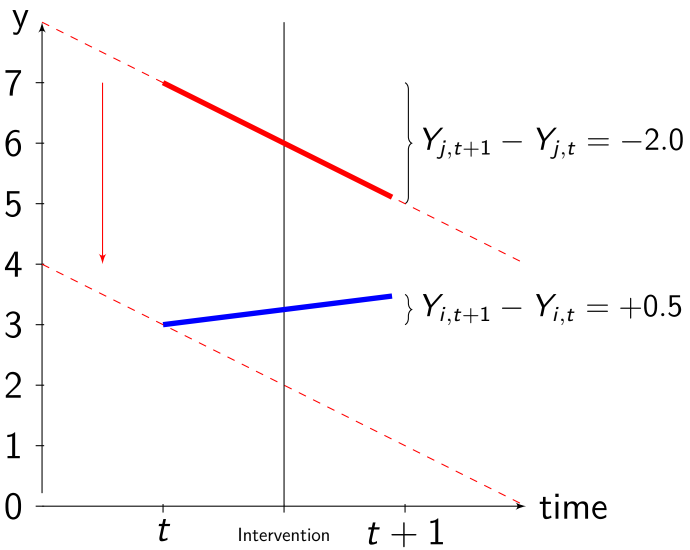
- Advantageous because variance for paired samples decreases as correlation between t_0 and t_1 observations increases

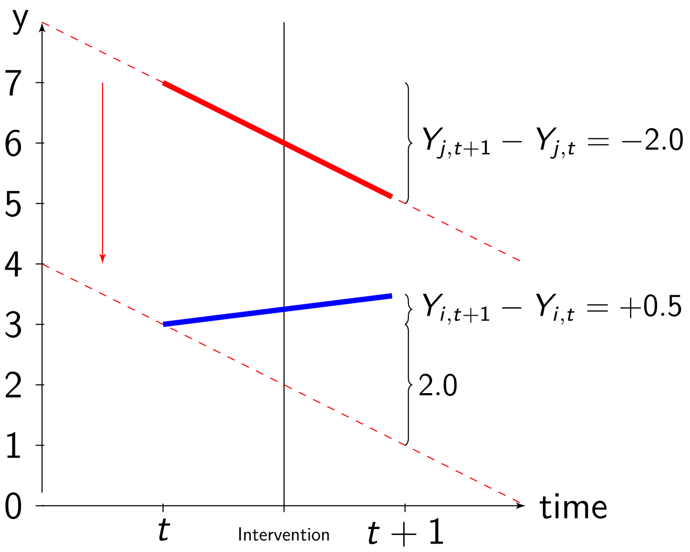

y
7
6
5
4
3
2
1
0
t
Intervention
t + 1
time

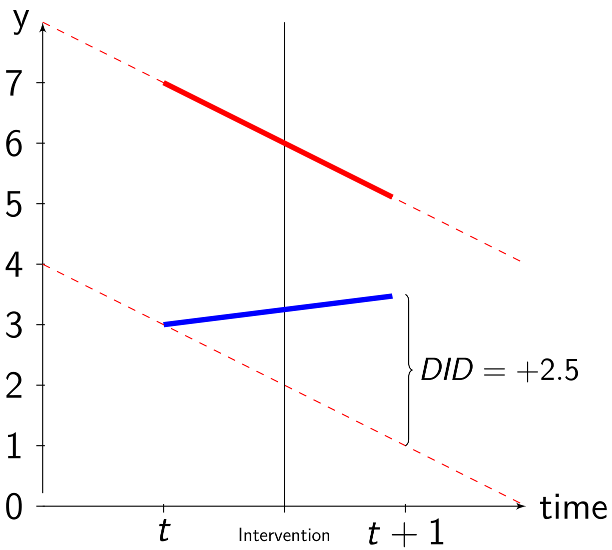

y
7
6
5
4
3
2
1
0
Treated
time
t
Intervention
t + 1


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


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


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


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


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

Threats to Validity

As soon as time comes into play, we have to worry about threats to validity.[1]

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

Threats to Validity

As soon as time comes into play, we have to worry about threats to validity.[1]

1. History (simultaneous cause)

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

Threats to Validity

As soon as time comes into play, we have to worry about threats to validity.[1]

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

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

Threats to Validity

As soon as time comes into play, we have to worry about threats to validity.[1]

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

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

Threats to Validity

As soon as time comes into play, we have to worry about threats to validity.[1]

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

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

Threats to Validity

As soon as time comes into play, we have to worry about threats to validity.[1]

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

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

Threats to Validity

As soon as time comes into play, we have to worry about threats to validity.[1]

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

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

III. Randomized Field Treatment

- Examples:

III. Randomized Field Treatment

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

III. Randomized Field Treatment

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

III. Randomized Field Treatment

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

III. Randomized Field Treatment

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

III. Randomized Field Treatment

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

Noncompliance

- Compliance is when individuals receive and accept the treatment to which they are assigned
- Noncompliance:
  "when subjects who were assigned to receive the treatment go untreated or when subjects assigned to the control group are treated" [2]
- This causes problems for our analysis because factors other than randomization explain why individuals receive their treatment
- Lots of methods for dealing with this, but the consequence is generally reduced power

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

Asymmetric Noncompliance

- Noncompliance *asymmetric* if only in one group
- We can ignore non-compliance and analyze the "intention to treat" effect, which will underestimate our effects because some people were not treated as assigned $ITT = \overline{Y}_1 - \overline{Y}_0$
- We can use "instrumental variables" to estimate the "local average treatment effect" (LATE) for those that complied with treatment: $LATE = \frac{ITT}{PercentCompliant}$
- We can ignore randomization and analyze data "as-treated", but this makes our study no longer an experiment

Local Average Treatment Effect

- IV estimate is *local* to the variation in X that is due to variation in D
- 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!

Two-Sided Noncompliance

- Two-sided noncompliance is more complex analytically
- Stronger assumptions are required to analyze it and we won't discus them here
- Best to try to develop a better design to avoid this rather than try to deal with the complexities of analyzing a broken design

IV. Treatment Encouragement

- Design:
  - T1: Encourage treatment
  - T2: Measure effects
- Examples:
  1. Albertson and Lawrence[3]

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

IV. Treatment Encouragement

- Design:
  - T1: Encourage treatment
  - T2: Measure effects
- Examples:
  1. Albertson and Lawrence[3]
- Issues

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

IV. Treatment Encouragement

- Design:
  - T1: Encourage treatment
  - T2: Measure effects
- Examples:
  1. Albertson and Lawrence[3]
- Issues
  - Nonresponse
  - Noncompliance

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

Treatment Noncompliance

Treatment Noncompliance

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

Questions?

Unrepresentative Samples

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

Unrepresentative Samples

- 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?

Unrepresentative Samples

- 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

Reweighting

- It may be possible to *reweight* convenience sample data to match a population
- Any method for this is "model-based" (rather than "design-based")
- Not widely used or evaluated (yet)
- All techniques build on the idea of stratification

Overview of Stratification

1. Define population
2. Construct a sampling frame
3. Identify variables we already know about units in the sampling frame
4. Stratify sampling frame based on these characteristics
5. Collect an SRS (of some size) within each stratum
6. Aggregate our results

Estimates from a stratified sample

- Within-strata estimates are calculated just like an SRS
- Within-strata variances are calculated just like an SRS

- Sample-level estimates are weighted averages of stratum-specific estimates
- Sample-level variances are weighted averages of strataum-specific variances

Post-Stratification

- Used to correct for nonresponse, coverage errors, and sampling errors

Post-Stratification

- Used to correct for nonresponse, coverage errors, and sampling errors
- Reweight sample data to match population distributions
    - Divide sample and population into strata
    - Weight units in each stratum so that the weighted sample stratum contains the same proportion of units as the population stratum does

Post-Stratification

- Used to correct for nonresponse, coverage errors, and sampling errors
- Reweight sample data to match population distributions
  - Divide sample and population into strata
  - Weight units in each stratum so that the weighted sample stratum contains the same proportion of units as the population stratum does
- There are numerous other related techniques

Post-Stratification: Example

- Imagine our sample ends up skewed on immigration status and gender relative to the population

Group	Pop.	Sample	Rep.	Weight
Native-born, Female	.45	.5		
Native-born, Male	.45	.4		
Immigrant, Female	.05	.07		
Immigrant, Male	.05	.03		

- PS weight is just $w_{ps} = N_l / n_l$

Post-Stratification: Example

- Imagine our sample ends up skewed on immigration status and gender relative to the population

Group	Pop.	Sample	Rep.	Weight
Native-born, Female	.45	.5	Over	
Native-born, Male	.45	.4	Under	
Immigrant, Female	.05	.07	Over	
Immigrant, Male	.05	.03	Under	

- PS weight is just $w_{ps} = N_l / n_l$

Post-Stratification: Example

- Imagine our sample ends up skewed on immigration status and gender relative to the population

Group	Pop.	Sample	Rep.	Weight
Native-born, Female	.45	.5	Over	0.900
Native-born, Male	.45	.4	Under	
Immigrant, Female	.05	.07	Over	
Immigrant, Male	.05	.03	Under	

- PS weight is just $w_{ps} = N_l / n_l$

Post-Stratification: Example

- Imagine our sample ends up skewed on immigration status and gender relative to the population

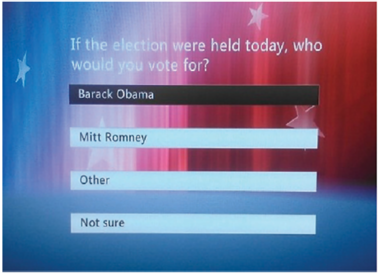
Group	Pop.	Sample	Rep.	Weight
Native-born, Female	.45	.5	Over	0.900
Native-born, Male	.45	.4	Under	1.125
Immigrant, Female	.05	.07	Over	0.714
Immigrant, Male	.05	.03	Under	1.667

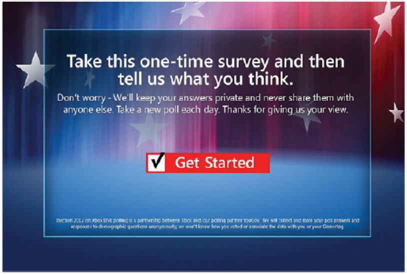
- PS weight is just $w_{ps} = N_l / n_l$

Post-Stratification

- This is the basis for inference in non-probability samples
  - *Demographic* representativeness
- Online panels will reweight sample based on age, sex, education, etc.
- Purely design-based surveys are increasingly rare

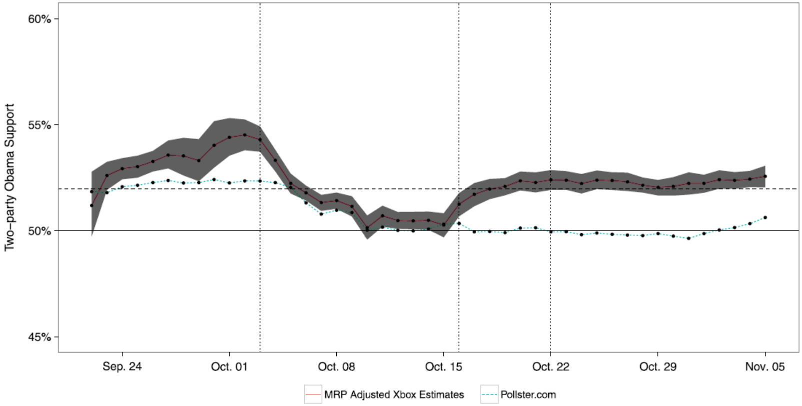
The Xbox Study



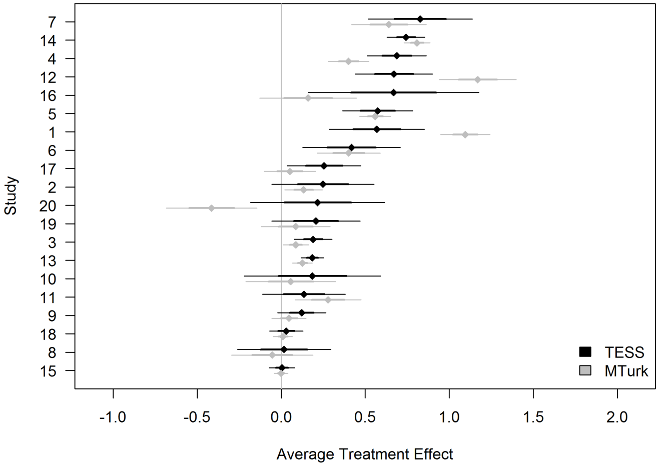


Wang et al. 2015. "Forecasting elections with non-representative polls." *International Journal of Forecasting*.

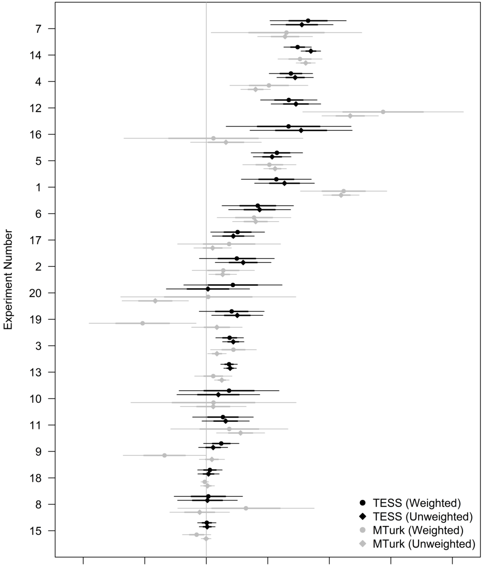
The Xbox Study



Wang et al. 2015. "Forecasting elections with non-representative polls." *International Journal of Forecasting*



Mullinix et al. In press. "The Generalizability of Survey Experiments."


Experiment Number
7
14
4
12
16
5
1
6
17
2
20
19
3
13
10
11
9
18
8
15
TESS (Weighted)
TESS (Unweighted)
MTurk (Weighted)
MTurk (Unweighted)

So does reweighting solve everything forever?

So does reweighting solve everything forever?

- Need well-defined target population
  - and detailed covariate data
  - and large stratum sizes

So does reweighting solve everything forever?

- Need well-defined target population
  - and detailed covariate data
  - and large stratum sizes
- Purely model-based, so only as good as the model
  - What unobservables might be hiding bias?
  - What reweighting might worse bias?

So does reweighting solve everything forever?

- Need well-defined target population
  - and detailed covariate data
  - and large stratum sizes
- Purely model-based, so only as good as the model
  - What unobservables might be hiding bias?
  - What reweighting might worse bias?
- Non-coverage is a potentially huge problem

So does reweighting solve everything forever?

- Need well-defined target population
  - and detailed covariate data
  - and large stratum sizes
- Purely model-based, so only as good as the model
  - What unobservables might be hiding bias?
  - What reweighting might worse bias?
- Non-coverage is a potentially huge problem
- Not well-tested on experimental data

Effect Mediation

- Sometimes we care about *why* an effect comes about (i.e., what is the mechanism, or mediator?)
- If we suspect this happens and we care about the mediation process, we should try to manipulate the treatment and the suspected mediator
- If we cannot manipulate the mediator, there is basically no credible way of estimating the "mediation effect" of the treatment group a given mediator
  - Have to assume exclusion restriction

SUTO Framework

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

Population

- Setting
- Units
- Treatments
- Outcomes

Your Study

- Setting
- Units
- Treatments
- Outcomes

Population

- Setting
- Units
- Treatments
- Outcomes

Your Study

- Setting
- Units
- Treatments
- Outcomes

In your study, how do these correspond?

Population

- Setting
- Units
- Treatments
- Outcomes

Your Study

- Setting
- Units
- Treatments
- Outcomes

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

Population

- Setting
- Units
- Treatments
- Outcomes

Your Study

- Setting
- Units
- Treatments
- Outcomes

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

Common Differences

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

Common Differences

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

Common Differences

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

Heterogeneity due to *S*ettings

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

Heterogeneity due to *S*ettings

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

Heterogeneity due to *S*ettings

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

Heterogeneity due to *S*ettings

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

Heterogeneity due to *S*ettings

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

Pretreatment Dynamics

"If the experiment explores a communication that regularly occurs in 'reality,' then reactions in the experiment might be contaminated by those 'regular' occurrences prior to the experiment." [4]

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

Pretreatment Dynamics

- Pretreatment is a feature of an experimental setting, treatment, and sample, wherein the effect of the treatment has already occurred[5]

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

Pretreatment Dynamics

- Pretreatment is a feature of an experimental setting, treatment, and sample, wherein the effect of the treatment has already occurred[5]
- Consequences:
  - Biased effect estimates

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

Pretreatment Dynamics

- Pretreatment is a feature of an experimental setting, treatment, and sample, wherein the effect of the treatment has already occurred[5]
- Consequences:
  - Biased effect estimates
- Mitigation:
  - Measure pretreatment
  - Avoid "pretreated" treatments or contexts
  - Study units not already treated
  - Theorize repeated effects

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

Questions?

Heterogeneity due to *U*nits

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

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

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

Block Randomization

Block Randomization

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

Block Randomization

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

Block Randomization

- Basic idea: randomization occurs within strata defined before treatment assignment
- CATE is estimate for each stratum; aggregated to SATE
- But...
  - 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, so works in panels but not cross-sectional designs
  - More precise SATE estimate

Questions?

Three Post-hoc Approaches

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

Three Post-hoc Approaches

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

Suggestive Evidence

We can never know $Var(TE_i)$!

Suggestive Evidence

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

- Quantile-quantile plots
- Equality of variance tests

Suggestive Evidence

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

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

Suggestive Evidence

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

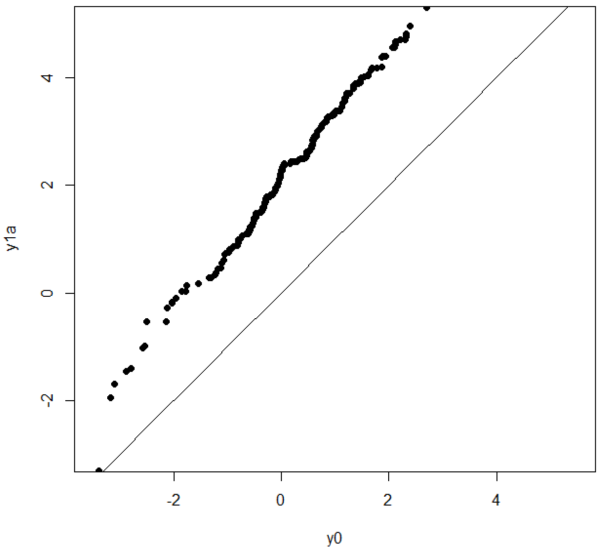
- Quantile-quantile plots
  - Compare the distribution of Y_0's to distribution of Y_1's
  - If homogeneity, a vertical shift in Y_1's
  - If heterogeneity, a slope $\neq 1$
- Equality of variance tests
  - If homogeneity, variance should be equal
  - If heterogeneity, variances should differ

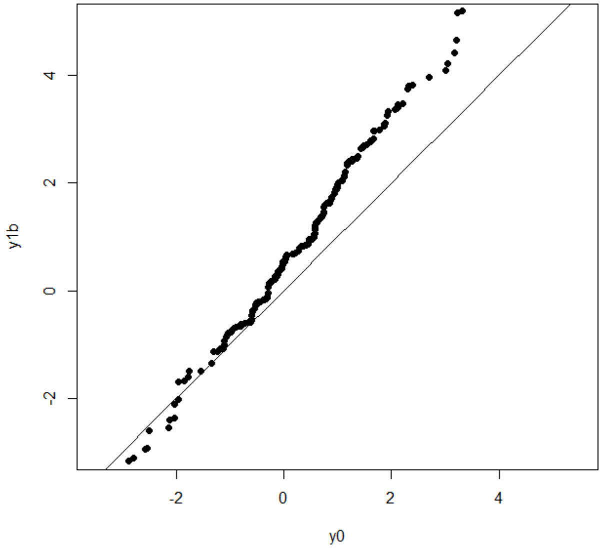
QQ Plots

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

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

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


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


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

Equality of Variance tests

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

        F test to compare two variances

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

Equality of Variance tests

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

        F test to compare two variances

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

Questions?

Regression Estimation

Aside: Regression Adjustment in Experiments, Generally

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

Scenario

Imagine two regression models. Which is correct?

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

Scenario

Imagine two regression models. Which is correct?

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

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

Treatment-Covariate Interactions

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

Treatment-Covariate Interactions

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

Treatment-Covariate Interactions

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

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

Treatment-Covariate Interactions

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

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

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

Let's work in Stata!

(Covariate-related effect heterogeneity)

BART

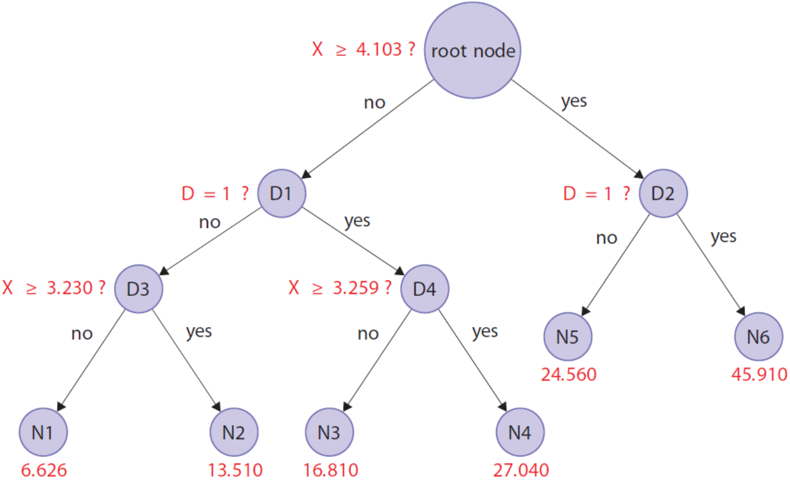
- Estimate CATEs in a fully automated fashion

BART

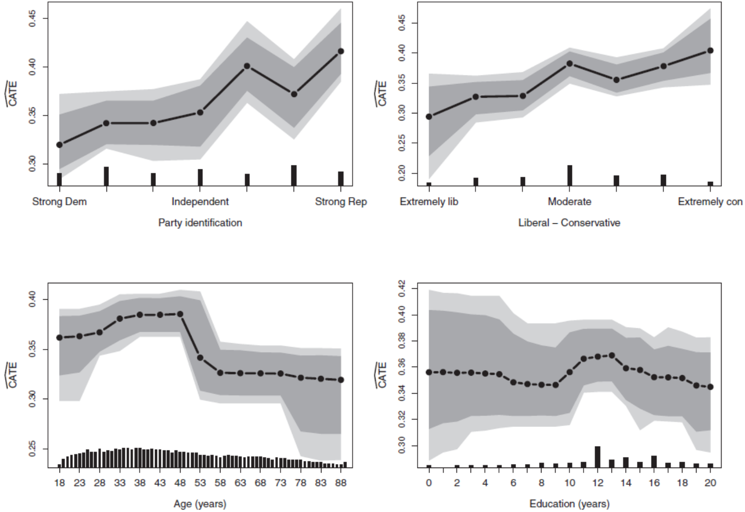
- Estimate CATEs in a fully automated fashion
- “Bayesian Additive Regression Trees”
  - Essentially an ensemble machine learning method

BART

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



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



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

Considerations

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

Considerations

Considerations

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

Considerations

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

Considerations

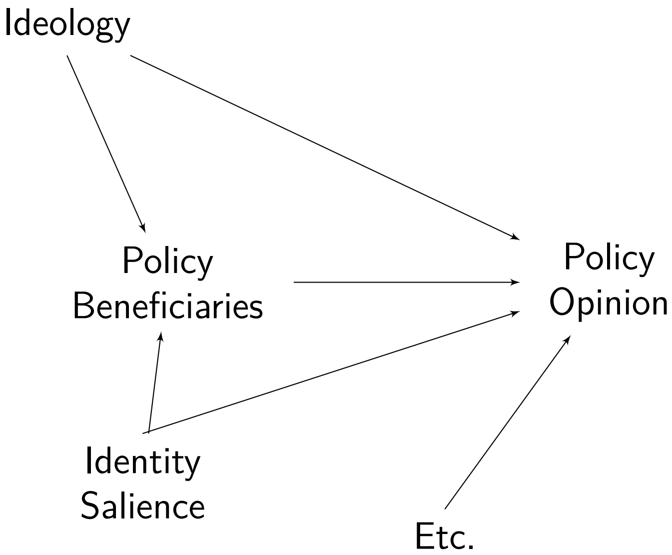
- Coefficients on moderators have no causal interpretation without further conditioning on observables
- Nearly unlimited potential moderators
  - First-order interactions with every covariate in dataset
  - Second-, third-order, etc. interactions
- Thus, multiple comparisons problem!
- Power (esp. if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

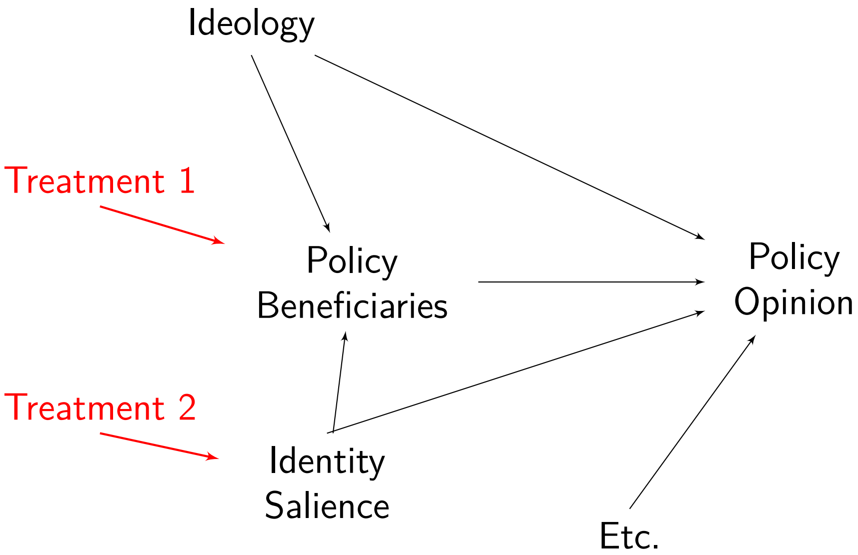
Simply: Manipulating the moderator variable is the best way to estimate a heterogeneous effect!

Why is this true?

Complex Designs

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


Ideology
Policy Beneficiaries
Policy Opinion
Identity Salience
Etc.


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

Ex. Question-as-treatment[6]

- 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?

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

Ex. Question-as-treatment[6]

- 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?

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

Ex. Question-as-treatment[6]

- 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?

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

Ex. Question-as-treatment[6]

- 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?

---

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

2x2 Factorial Design

Condition	
Educ. for Minorities	Y_1
Schools	Y_0

2x2 Factorial Design

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

Two ways to *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

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

Probably obvious, but...

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

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

Considerations

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

Considerations

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

Questions?

Treatment Preferences/Self-Selection

Bennett and Iyengar:[7]

> *manipulational control actually weakens the ability to generalize to the real world where exposure to stimuli is typically voluntary. Accordingly, it is important that experimental researchers use designs that combine manipulation with self-selection of exposure.*

---

[7] p.724 from Bennett & Iyengar. 2008. "A new era of minimal effects? The changing foundations of political communication." *Journal of Communication* 58(4): 707–31.

Hovland: [8]

> *It should be possible to assess what demographic and personality factors predispose one to expose oneself to particular communications and then to utilize experimental and control groups having these characteristics. Under some circumstances the evaluation could be made on only those who select themselves, with both experimental and control groups coming from the self-selected audience.*

[8] p.16 from Hovland. 1959. "Reconciling conflicting results derived from experimental and survey studies of attitude change." *American Psychologist* 14(1): 8–17.

Treatment Preferences I

- Experiments are about inferring effect of X on Y
- Respondents may have preferences over whether they are treated or untreated[9]
- Origins of this discussion are in the medical literature[10]
- Closely related to the notion of placebo effects

[9] Rucker. 1989. "A Two-Stage Trial Design for Testing Treatment, Self-Selection, and Treatment Preference Effects." *Statistics in Medicine* 8: 477–485.

[10] Swift & Callahan. 2009. "The Impact of Client Treatment Preferences on Outcome: A Meta-Analysis." *Journal of Clinical Psychology* 65(4): 368–381.

Treatment Preferences I

- Treatment preferences may be an important factor in:
    - Compliance
    - Effect heterogeneity

Treatment Preferences I

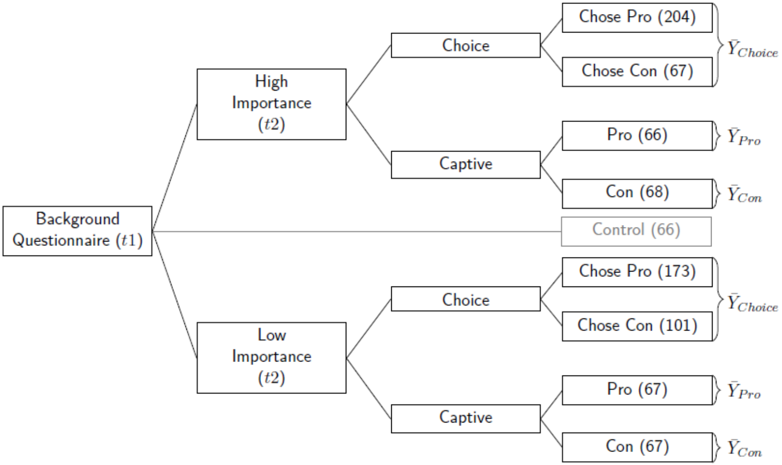
- Treatment preferences may be an important factor in:
  - Compliance
  - Effect heterogeneity
- Depending on your treatments, you may want to measure preferences

Treatment Preferences I

- Treatment preferences may be an important factor in:
  - Compliance
  - Effect heterogeneity
- Depending on your treatments, you may want to measure preferences
  1. Stated preference measures

Treatment Preferences I

- Treatment preferences may be an important factor in:
    - Compliance
    - Effect heterogeneity
- Depending on your treatments, you may want to measure preferences
    1. Stated preference measures
    2. Designs that reveal preferences



Leeper. 2017. "How Does Treatment Self-Selection Affect Inferences About Political Communication?" *Journal of Experimental Political Science*: In Press. Available at `http://thomasleeper.com/research.html`

Analyzing 3-Group Preference Trials[11]

1. SATE: $\bar{Y}_T - \bar{Y}_C$
2. CATE (Prefer T): $\dfrac{\bar{Y}_{Choice} - \bar{Y}_C}{\hat{\alpha}}$
3. CATE (Prefer C): $\dfrac{\bar{Y}_T - \bar{Y}_{Choice}}{1 - \hat{\alpha}}$

Note: $\alpha = Pr(T|Choice)$

[11] GK2011 Package for R. `https://cran.r-project.org/package=GK2011`

Questions?

Attention and Satisficing

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

Attention and Satisficing

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:

1. Noncompliance (discussed earlier)
2. Survey Satisficing
3. Apparent Inattention

Substantive Manipulation Check

- Two common approaches:
  - Information recall or understanding
  - Measure level of manipulated treatment variable
- Risky to remove cases based on this because it is a form of conditioning on post-treatment variables
- May be useful to consider either a mediator of effects

Attention Checking

- Online mode invites satisficing
- Attention checking can help, but is imperfect

Apparent Satisficing

- Filter out respondents based on response behavior
- Some common measures:
  - "Straightlining"
  - Non-differentiation
  - Acquiescence
  - Nonresponse
  - DK responding
  - Speeding
- Difficult to detect
- Difficult to 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

Attention Checking

In summary...

- Attention checking can be useful
- Lots of options
- No obvious best metric
- Can be analytically consequential

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

1. Keep them
2. Throw them away

Best Practice: Protocol

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

When are you excluding participants?

Pre-Treatment	Post-Treatment

When are you excluding participants?

Pre-Treatment

- Satisficing behaviors

Post-Treatment

When are you excluding participants?

Pre-Treatment

- Satisficing behaviors
- Inattention

Post-Treatment

When are you excluding participants?

Pre-Treatment

- Satisficing behaviors
- Inattention
- Covariate-based selection

Post-Treatment

When are you excluding participants?

Pre-Treatment

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

Post-Treatment

When are you excluding participants?

Pre-Treatment

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

Post-Treatment

- Speeding on treatment

When are you excluding participants?

Pre-Treatment

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

Post-Treatment

- Speeding on treatment
- "Failing" a manipulation check

When are you excluding participants?

Pre-Treatment

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

Post-Treatment

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

Pre-Treatment Exclusion

- This is totally fine from a causal inference perspective

Pre-Treatment Exclusion

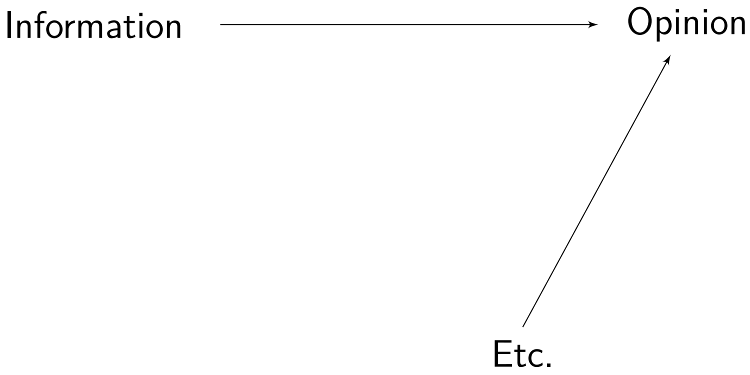
- This is totally fine from a causal inference perspective
- Advantages:
  - Focused on engaged respondents
  - Likely increase impact of treatment

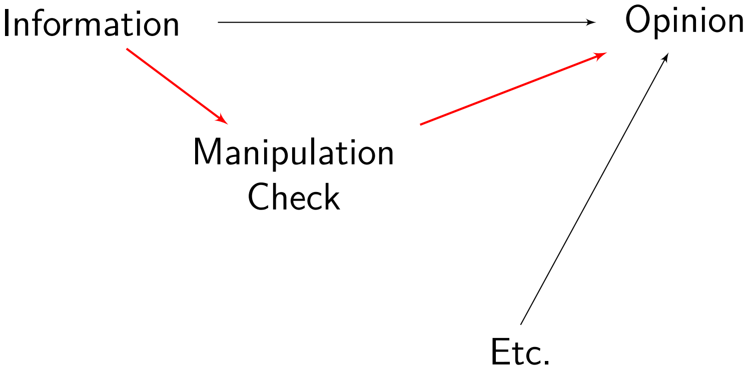
Pre-Treatment Exclusion

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

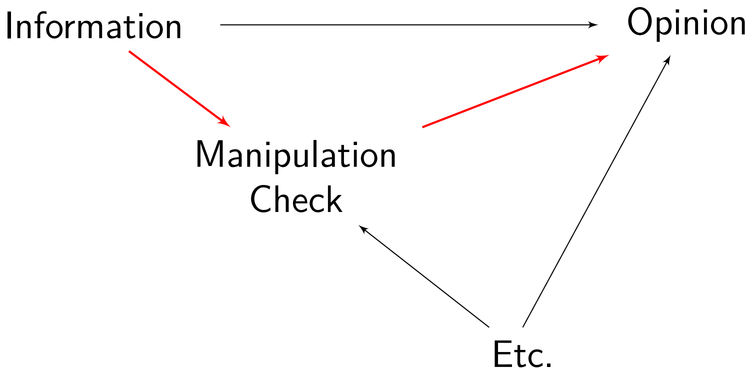
Post-Treatment Exclusion

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


Information
Opinion
Etc.



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



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

Post-Treatment Exclusion

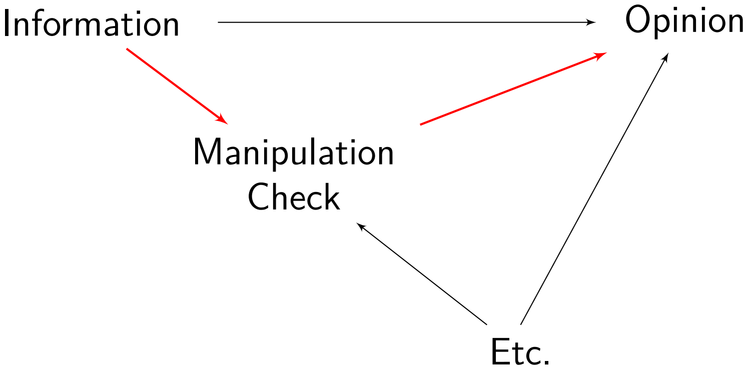
- Any post-treatment exclusion is problematic and should be avoided

Post-Treatment Exclusion

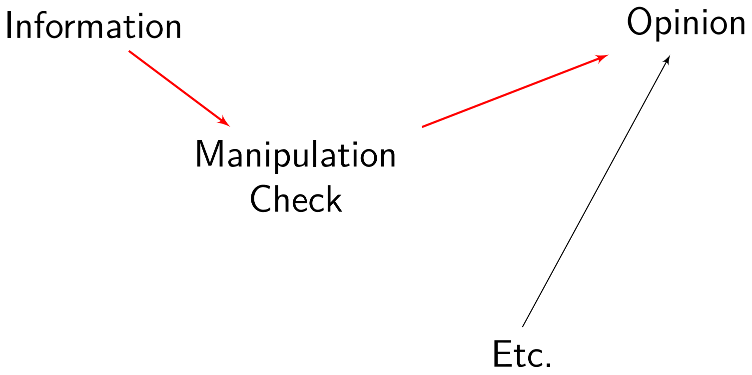
- Any post-treatment exclusion is problematic and should be avoided
- Can estimate a LATE
    - Interpretation: Effect of manipulation check among those whose value of the check can be changed by the treatment manipulation

Post-Treatment Exclusion

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



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


Information
Opinion
Manipulation
Check
Etc.

Post-Treatment Exclusion

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

Post-Treatment Exclusion

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

Questions?

Heterogeneity due to *T*reatments

- We should expect this! Why?

Heterogeneity due to *T*reatments

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

Conjoint Designs I

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

Conjoint Designs I

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

Conjoint Designs I

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

Conjoint Designs II

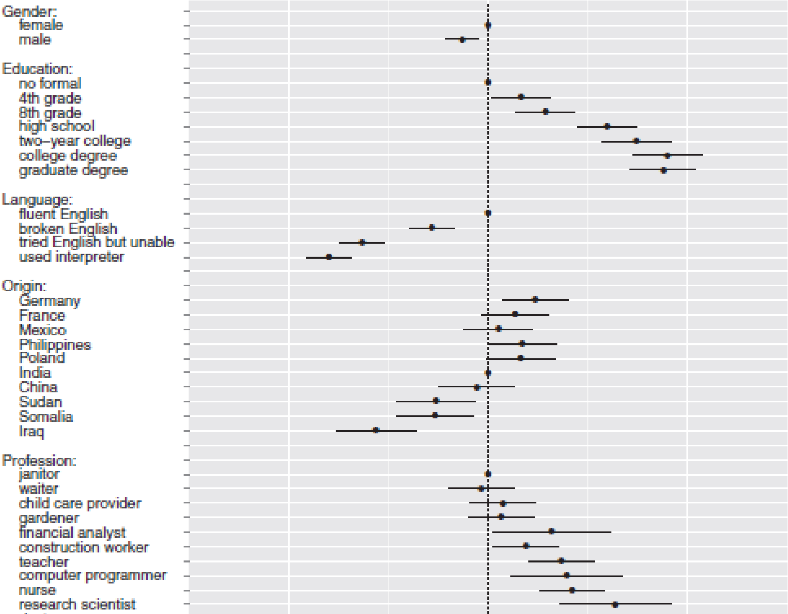
Why is this useful?

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

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

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

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


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

Conjoint Designs III

Conjoint Designs III

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

Conjoint Designs III

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

Replication

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

Replication

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

Replication

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

Questions?

Heterogeneity due to *O*utcomes

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

Heterogeneity due to *O*utcomes

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

Heterogeneity due to *O*utcomes

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

Behavioural measures

Some behaviours that can be directly measured through survey questionnaires.

Behavioural measures

Some behaviours that can be directly measured through survey questionnaires.

Three broad categories:

Behavioural measures

Some behaviours that can be directly measured through survey questionnaires.

Three broad categories:

1. Behavioural measures that provide survey paradata

Behavioural measures

Some behaviours that can be directly measured through survey questionnaires.

Three broad categories:

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

Behavioural measures

Some behaviours that can be directly measured through survey questionnaires.

Three broad categories:

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

Behavioural Measures for Paradata

Why?

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

Behavioural Measures for Paradata

Why?

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

What?

Behavioural Measures for Paradata

Why?

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

What?

- Nonresponse

Behavioural Measures for Paradata

Why?

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

What?

- Nonresponse
- Response latencies

Behavioural Measures for Paradata

Why?

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

What?

- Nonresponse
- Response latencies
- Reading times

Behavioural Measures for Paradata

Why?

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

What?

- Nonresponse
- Response latencies
- Reading times
- Answer switching

Behavioural Measures for Paradata

Why?

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

What?

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

Behavioural Measures for Paradata

Why?

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

What?

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

Behavioural Measures for Paradata

Why?

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

What?

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

Behavioural Measures for Attitudes

Why?

- Attitudinal self-reports might be "cheap talk"

Behavioural Measures for Attitudes

Why?

- Attitudinal self-reports might be "cheap talk"

What?

Behavioural Measures for Attitudes

Why?

- Attitudinal self-reports might be "cheap talk"

What?

- Implicit Association Test

Behavioural Measures for Attitudes

Why?

- Attitudinal self-reports might be “cheap talk”

What?

- Implicit Association Test
- Incentivized Survey questions

Behavioural Measures for Behaviour

Why?

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

Behavioural Measures for Behaviour

Why?

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

What?

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

Example 1: Active Information Choice

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

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

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

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

Example 1: Active Information Choice

- "Followed link" identification[12]

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

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

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

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

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

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

Example 1: Active Information Choice

- "Followed link" identification[12]

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

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

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

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

Example 1: Active Information Choice

- "Followed link" identification[12]
- Information boards[13]

---

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

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

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

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

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

Example 1: Active Information Choice

- "Followed link" identification[12]
- Information boards[13]

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

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

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

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

Example 1: Active Information Choice

- "Followed link" identification[12]
- Information boards[13]
- Video choice[14]

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

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

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

[15] https://dpte.polisci.uiowa.edu/dpte/

Example 1: Active Information Choice

- "Followed link" identification[12]
- Information boards[13]
- Video choice[14]
- Dynamic Process Tracing Environment [15]

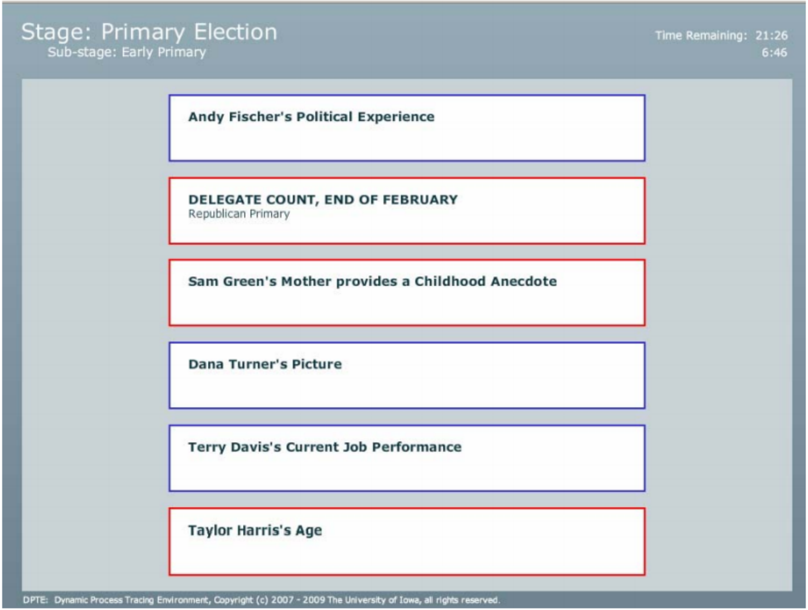
---

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

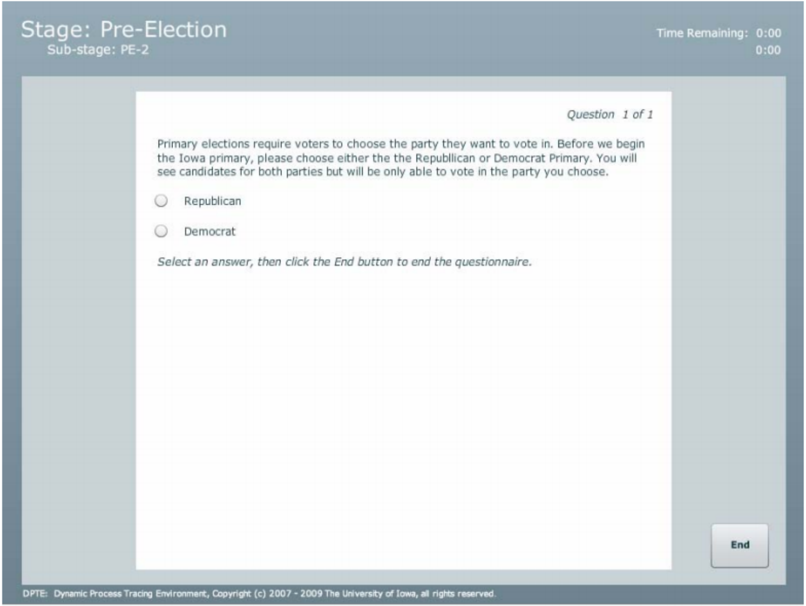
[13] Leeper, TJ. 2014. "The Informational Basis for Mass Polarization." *Public Opinion Quarterly* 78(1): 27–46. doi:10.1093/poq/nft045

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

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


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


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



Stage: Pre-Election
Sub-stage: PE-2

Time Remaining: 0:00
0:00

Question 1 of 1

Primary elections require voters to choose the party they want to vote in. Before we begin the Iowa primary, please choose either the the Republlican or Democrat Primary. You will see candidates for both parties but will be only able to vote in the party you choose.

- Republican
- Democrat

Select an answer, then click the End button to end the questionnaire.

End

Example 2: Sign-up/Enrolment

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

- Entering an email address to receive information or join a mailing list [16] [17]
- Signing up for an appointment or further interaction

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

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

Example 3: Incentivised Survey Questions

Definitions:

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

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

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

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

Example 3: Incentivised Survey Questions

Definitions:

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

Example 3: Incentivised Survey Questions

Definitions:

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

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

Example 4: Purchasing Decisions

Common ways to study purchasing behaviour include:

Example 4: Purchasing Decisions

Common ways to study purchasing behaviour include:

- Direct attitudinal questions

Example 4: Purchasing Decisions

Common ways to study purchasing behaviour include:

- Direct attitudinal questions
- Retrospective and prospective self-reports

Example 4: Purchasing Decisions

Common ways to study purchasing behaviour include:

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

Example 4: Purchasing Decisions

Common ways to study purchasing behaviour include:

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

Another way is embedding a purchase in a survey.[18]

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

Source: Wikimedia Commons (Sun Ladder, KMJ)

Example 5: Donations

- Miller and Krosnick[19] asked for charitable donations via cheque directly as part of a paper-and-pencil survey

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

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

Example 5: Donations

- Miller and Krosnick[19] asked for charitable donations via cheque directly as part of a paper-and-pencil survey
- Klar and Piston[20] offered respondents a survey incentive up-front for participation and then later offered them a chance to donate (a portion of payment) to a charity

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

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

Example 6: Web Tracking Data

1. Active installation of a tracking app, such as YouGov Pulse[21] [22]
2. Post-hoc collection of web history files using something like Web Historian [23]

---

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

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

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

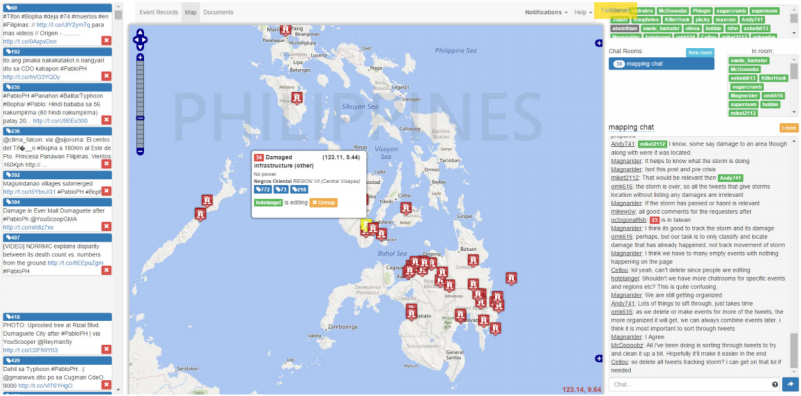
Other Possibilities

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

Other Possibilities

- Coordination tasks
  - Synchronous group tasks[24]
  - Game play
  - Simulations

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



Event Records
Map
Documents
Notifications
Help
PHILIPPINES
Philippine Sea
Sibuyan Sea
Visayan Sea
Bohol Sea
Damaged infrastructure (other)
(123.11, 9.44)
No power.
Negros Oriental REGION VII (Central Visayas)
hotelangel is editing
Unmap
123.14, 9.64
Chat Rooms:
New room
In room:
mapping chat
Leave
Andy741: I know, some say damage to an area though along with were it was located.
Magnarider: It helps to know what the storm is doing
Magnarider: Isnt this post and pre crisis
miket2112: That would be relevant then
omk616: the storm is over, so all the tweets that give storms location without listing any damages are irrelevant
Magnarider: If the storm has passed or hasnt is relevant
mikew0w: all good comments for the requesters after
octogonalfish: is in taiwan
Magnarider: I think its good to track the storm and its damage
omk616: perhaps, but our task is to only classify and locate damage that has already happened, not track movement of storm
Magnarider: I think we have to many empty events with nothing happening on the page
Cellou: lol yeah, can't delete since people are editing
hotelangel: Shouldn't we have more chatrooms for specific events and regions etc? This is quite confusing.
Magnarider: We are still getting organized
Andy741: Lots of things to sift through, just takes time
omk616: as we delete or make events for more of the tweets, the more organized it will get, we can always combine events later. i think it is most important to sort through tweets
Magnarider: I Agree
McDoooobz: All I've been doing is sorting through tweets to try and clean it up a bit. Hopefully it'll make it easier in the end
Cellou: so delete all tweets tracking storm? i can get on that lol if needed
Chat...

Other Possibilities

- Coordination tasks
  - Synchronous group tasks[24]
  - Game play
  - Simulations

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

Other Possibilities

- Coordination tasks
  - Synchronous group tasks[24]
  - Game play
  - Simulations
- Offering incentives to perform future behaviour (tracked elsewhere)

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

Other Possibilities

- Coordination tasks
  - Synchronous group tasks[24]
  - Game play
  - Simulations
- Offering incentives to perform future behaviour (tracked elsewhere)
- OAuth/API integrations w/ other platforms
  - Merging website usage data w/ survey data
  - Treating website sign-up or usage as behavioural outcomes
  - Linking with smartphone metadata

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

Some principles for survey measures of behaviour

Some principles for survey measures of behaviour

1. Know why you are collecting a behavioural measure!

Some principles for survey measures of behaviour

1. Know why you are collecting a behavioural measure!
2. Know whether you are studying a past, present, or future behaviour.

Some principles for survey measures of behaviour

1. Know why you are collecting a behavioural measure!
2. Know whether you are studying a past, present, or future behaviour.
3. Be creative! Recognise possibilities and limitations of any given survey mode.

Some principles for survey measures of behaviour

1. Know why you are collecting a behavioural measure!
2. Know whether you are studying a past, present, or future behaviour.
3. Be creative! Recognise possibilities and limitations of any given survey mode.
4. Validate, validate, validate!

Activity!

With a partner, brainstorm how one or more these behavioural measures might be applied to a survey experiment (either as outcome, treatment, covariate, or behavioural check) relevant to your own work or your organisation.

SUTO Punchline 1: Replication!

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

SUTO Punchline 1: Replication!

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

SUTO Punchline 1: Replication!

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

SUTO Punchline 2: What do you want to know?

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

SUTO Punchline 2: What do you want to know?

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

SUTO Punchline 2: What do you want to know?

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

Questions?

Recruitment Considerations

- Recruitment
  - Sampling
  - Opt-in
  - A mix of each
- Incentives
- Frequency of participation
  - MTurk panelists do 100+ studies per month
  - YouGov panelists do nearly as many
- "Profile" variables
- Quotas, post-stratification, weighting
- Respondent "quality"

Professional Panels

- Big players: SSI, YouGov, GfK, TNS/Gallup
- Online panels of respondents
- Respondents participate for incentives
- Study costs are negotiated
  - Sample size
  - Study length (number of survey items)
  - Targeting
  - Timing

KnowledgeNetworks versus YouGov

- Big debate in early 2000s about online panels
  - KN used ABS to build a representative panel
  - YouGov created an opt-in panel; used “sample matching”

KnowledgeNetworks versus YouGov

- Big debate in early 2000s about online panels
  - KN used ABS to build a representative panel
  - YouGov created an opt-in panel; used "sample matching"
- YouGov's process:
  - Randomly sample from a list
  - Match each sampled individual to someone in their opt-in panel
  - Survey the matched individuals

KnowledgeNetworks versus YouGov

- Big debate in early 2000s about online panels
  - KN used ABS to build a representative panel
  - YouGov created an opt-in panel; used "sample matching"
- YouGov's process:
  - Randomly sample from a list
  - Match each sampled individual to someone in their opt-in panel
  - Survey the matched individuals
- Evidence inconclusive but many think KN approach is better

Opt-in (Crowdsourcing) Sites

- Not exactly a panel (fully opt-in)
- Incentivized participation

Opt-in (Crowdsourcing) Sites

- Not exactly a panel (fully opt-in)
- Incentivized participation
- Prominent examples
  - MTurk
  - Crowdflower
  - Microworkers
  - Prolific Academic
  - Google Surveys

“River Sampling”

- Not using an existing subject pool
  - Link sharing or posting on websites
  - Using email list
  - Online advertising (Google, Facebook)

"River Sampling"

- Not using an existing subject pool
  - Link sharing or posting on websites
  - Using email list
  - Online advertising (Google, Facebook)
- My advice: don't do this unless you have no other choice!

Custom Panels

- Creating your own panel is great
  - Carefully sample on specific characteristics
  - Organize repeated interviewing or interaction
- Lots of additional issues
  - Attrition
  - Compensation
  - Panel Conditioning
- See Callegaro et al. 2014. *Online Panel Research: A Data Quality Perspective*. Wiley.

My Advice, Elaborated

- Only work with populations where each unit is uniquely identifiable

My Advice, Elaborated

- Only work with populations where each unit is uniquely identifiable
- Without this, you risk many things:
    - Ambiguous eligibility
    - Retakes, treatment crossover
    - No way to evaluate response rates/bias

My Advice, Elaborated

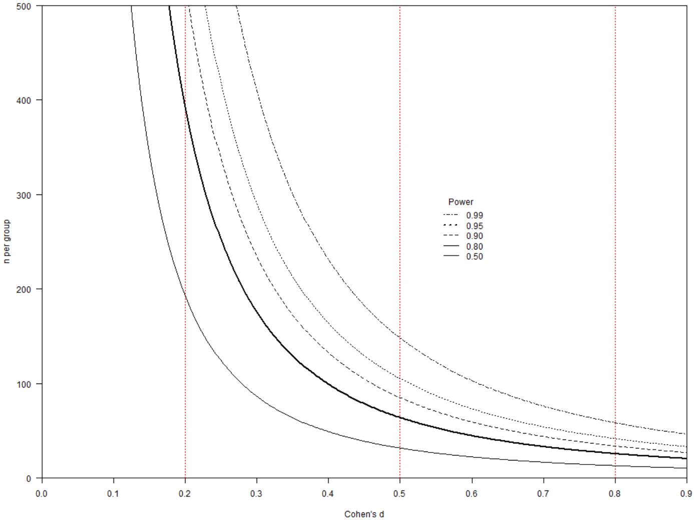
- Only work with populations where each unit is uniquely identifiable
- Without this, you risk many things:
  - Ambiguous eligibility
  - Retakes, treatment crossover
  - No way to evaluate response rates/bias
- Know something about your sample
  - How does it differ from your target of inference?
  - What theories or evidence would suggest those differences should matter?
  - What can you do to adjust or control for those *consequential* differences?

Measure, Measure, Measure

The only way to evaluate a sample is to know something about it.

The best way to convince reviewers is to rule out irrelevancies.

Don't forget statistical power. . .



And don't forget costs, either!

From one of my studies:

Sample	Cost	n	Cost/participant
National	$13200	593	$22.26
Exit Poll	$3000	741	$4.05
Students	$0	299	$0
Staff	$1280	128	$10.00
MTurk	$550	1024	$0.54
Ads	$636	80	$7.95

Quiz time!

Compliance

1 What is compliance?

Compliance

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

Balance testing

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

Balance testing

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

Balance testing

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

Nonresponse and Attrition

1. Do we care about outcome nonresponse in experiments?

Nonresponse and Attrition

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

Manipulation checks

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

Manipulation checks

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

Null effects

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

Null effects

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

Null effects

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

Effect heterogeneity

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

Effect heterogeneity

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

Representativeness

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

Representativeness

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

Peer Review

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

Peer Review

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

Questions?

Presentations!

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!

Experimental Methods

- Druckman et al. 2011. *Cambridge Handbook of Experimental Political Science*. Cambridge.
- Gerber and Green. 2011. *Field Experiments*. W.W. Norton.
- Mutz. 2011. *Population-Based Survey Experiments*. Princeton.
- Shadish et al. 2001. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Houghton Mifflin.

Survey Methods

- Groves et al. 2008. *Survey Methodology*. 2nd Edition. Wiley.
- Lohr. 2010. *Sampling: Design and Analysis*. 2nd Edition. Cengage.

Online Surveys

- Callegaro et al. 2015. *Web Survey Methodology*. Sage.
- Callegaro et al. 2015. *Online Panel Research: A Data Quality Perspective*. Wiley.