Session III
External Validity

Thomas J. Leeper

Government Department
London School of Economics and Political Science

Reminder: Friday

If you're interested in presenting an idea for a survey experiment on Friday, let me know in person or via email.

Review

What techniques can we use to assess whether a treatment manipulated what we wanted it to and did not manipulate what we didn't want it to?

Review

What are some of the available paradigms for implementing a survey experiment?

Review

What is an experimental protocol document and why is it useful?

1. External Validity of a Sample
   - Design-based
   - Model-based

2. Other Notions of External Validity

3. Other Survey Experimental Designs

1. External Validity of a Sample
   - Design-based
   - Model-based
2. Other Notions of External Validity
3. Other Survey Experimental Designs

Think–Pair–Share

Consider the following question:
What makes an experiment (or any research study) generalizable? What does it mean for a study's results to "generalize"?

1. Write or think to yourself for 90 seconds
2. Then, discuss with the person next to you

"The Gold Standard"

a population-based experiment uses survey sampling methods to produce a collection of experimental subjects that is representative of the target population of interest for a particular theory . . . the population represented by the sample should be representative of the population to which the researcher intends to extend his or her findings. In population-based experiments, experimental subjects are randomly assigned to conditions by the researcher

p2. from Mutz, Diana. 2011. *Population-Based Survey Experiments*. Princeton University Press.

Surveys Start with an Inference Population

- We want to speak to a population
- But what population is it?

Surveys Start with an Inference Population

- We want to speak to a population
- But what population is it?
    - A national population?

Surveys Start with an Inference Population

- We want to speak to a population
- But what population is it?
    - A national population?
    - Adults in Western, industrialized democracies?

Surveys Start with an Inference Population

- We want to speak to a population
- But what population is it?
  - A national population?
  - Adults in Western, industrialized democracies?
  - All human beings?

Surveys Start with an Inference Population

- We want to speak to a population
- But what population is it?
  - A national population?
  - Adults in Western, industrialized democracies?
  - All human beings?
- This is rarely specified, but is important when we think about whether a sample is appropriate

A Hypothetical Census

- Advantages

- Disadvantages

A Hypothetical Census

- Advantages
  - Perfectly representative
  - Sample statistics are population parameters
- Disadvantages

A Hypothetical Census

- Advantages
  - Perfectly representative
  - Sample statistics are population parameters
- Disadvantages
  - Costs
  - Feasibility
  - Need

1. External Validity of a Sample
   - Design-based
   - Model-based
2. Other Notions of External Validity
3. Other Survey Experimental Designs

Sampling Considerations...

- Design-based survey samples all work from the premise of each unit having a *known*, *non-zero* probability of being sampled
  - SRS is representative per se
  - Non-self-weighting samples representative when weighted

Sampling Considerations...

- Design-based survey samples all work from the premise of each unit having a *known, non-zero* probability of being sampled
  - SRS is representative per se
  - Non-self-weighting samples representative when weighted
- Random sampling ensures that samples are, *in expectation*, representative of the population *in all respects*
  - Demographics
  - Covariances
  - Potential outcomes

Representativeness

What does it mean for a sample to be representative?

Representativeness

What does it mean for a sample to be representative?

- Census?

Representativeness

What does it mean for a sample to be representative?

- Census?
- Probability-based sampling?

Representativeness

What does it mean for a sample to be representative?

- Census?
- Probability-based sampling?
- Quota fulfillment?

Representativeness

What does it mean for a sample to be representative?

- Census?
- Probability-based sampling?
- Quota fulfillment?
- Others?

Representativeness

What does it mean for a sample to be representative?

- Census?
- Probability-based sampling?
- Quota fulfillment?
- Others?

Which of these matter?

Combining Probability Sampling and Experimental Design

- Sample is representative of population in every respect (in expectation)
- Sample Average Treatment Effect (SATE) is the average of the sample's individual-level treatment effects
  - Unbiased estimate of PATE
  - Not necessarily any unit's individual treatment effect
  - Blocking might reduce variance
  - Optimized for estimating SATE

Credibility of all of this is based on *design* only

Credibility of all of this is based on *design* only

Sampling aspect only works in a world of perfect coverage and no response bias

My View

100% design-based inference does not exist!

My View

100% design-based inference does not exist!

- All survey designs involve reweighting adjustments

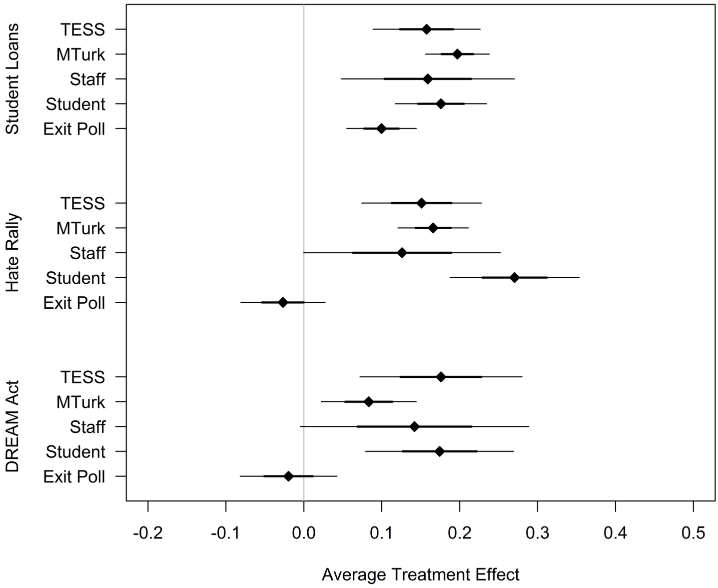
My View

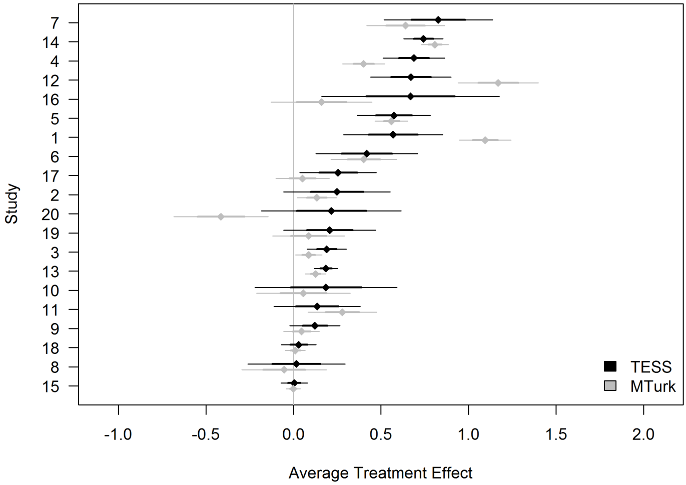
100% design-based inference does not exist!

- All survey designs involve reweighting adjustments
- Representativeness is a more complex issue than demographic comparisons

	GfK	Poll	Student	Staff	MTurk	Ads
Dem. (%)	51.3	86.1	75.7	66.4	62.1	72.1
Rep. (%)	46.0	7.7	17.8	16.4	20.3	14.7
Lib. (%)	27.8	75.4	68.5	62.7	60.4	66.2
Con. (%)	35.3	9.4	14.7	19.8	19.1	17.7
Fem. (%)	51.1	60.8	56.4	50.8	41.7	65.3
White (%)	77.9	67.6	62.9	60.2	76.0	53.8
Age	49.4	40-49	18-24	25-34	25-34	25-34
Interest	2.8	3.5	3.2	2.8	2.7	3.0
N	593	741	299	128	1024	80

Mullinix et al. 2015. "The Generalizability of Survey Experiments." *Journal of Experimental Political Science*.


Student Loans
TESS
MTurk
Staff
Student
Exit Poll
Hate Rally
TESS
MTurk
Staff
Student
Exit Poll
DREAM Act
TESS
MTurk
Staff
Student
Exit Poll
-0.2
-0.1
0.0
0.1
0.2
0.3
0.4
0.5
Average Treatment Effect



Study
7
14
4
12
16
5
1
6
17
2
20
19
3
13
10
11
9
18
8
15
TESS
MTurk
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
Average Treatment Effect

My View

100% design-based inference does not exist!

- All survey designs involve reweighting adjustments
- Representativeness is a more complex issue than demographic comparisons

My View

100% design-based inference does not exist!

- All survey designs involve reweighting adjustments
- Representativeness is a more complex issue than demographic comparisons
- Randomization gives us clear causal inference about a *local* effect
  - Sacrifice representativeness for causal inference
  - Try to figure nature of the *localness*

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

Shadish, Cook, and Campbell tell us to think about:

- Surface similarities
- Ruling out irrelevancies
- Making discriminations
- Interpolation/extrapolation

Common differences

Shadish, Cook, and Campbell tell us to think about:

- Surface similarities
- **Ruling out irrelevancies**
- **Making discriminations**
- Interpolation/extrapolation

Focus on effect heterogeneity

Focus on effect heterogeneity

- **Constant effects**: TE_i and Y_{i0} are same for all observations

Focus on effect heterogeneity

- **Constant effects**: TE_i and Y_{i0} are same for all observations
- **Homogeneous effects**: TE_i is same for all observations

Focus on effect heterogeneity

- **Constant effects**: TE_i and Y_{i0} are same for all observations
- **Homogeneous effects**: TE_i is same for all observations
- **Heterogeneous effects**: TE_i is different for all observations

Focus on effect heterogeneity

- Think about and make an evidence-based argument for why you think there are (or are not) heterogeneous effects

Focus on effect heterogeneity

- Think about and make an evidence-based argument for why you think there are (or are not) heterogeneous effects
- If you think there is heterogeneity, then we probably do not care about the SATE anyway

Focus on effect heterogeneity

- Think about and make an evidence-based argument for why you think there are (or are not) heterogeneous effects
- If you think there is heterogeneity, then we probably do not care about the SATE anyway
- Conditional Average Treatment Effect: $E[Y_{1i}|X=1,Z=z]-E[Y_{0i}|X=0,Z=z]$

1 External Validity of a Sample

- Design-based
- Model-based

2 Other Notions of External Validity

3 Other Survey Experimental Designs

Stratification/Blocking

As soon as we care about heterogeneous effects, it makes sense to stratify and block on factors that might moderate the treatment effect.

Stratification/Blocking

As soon as we care about heterogeneous effects, it makes sense to stratify and block on factors that might moderate the treatment effect.

As soon as we identify all sources of heterogeneity, it doesn't matter what sample we use because effects are *by definition* homogeneous within such strata.

Stratification/Blocking

As soon as we care about heterogeneous effects, it makes sense to stratify and block on factors that might moderate the treatment effect.

As soon as we identify all sources of heterogeneity, it doesn't matter what sample we use because effects are *by definition* homogeneous within such strata.

But, we never know when we've reached that point!

If we acknowledge and start thinking about effect heterogeneity, does this mean we can use any convenient group of participants as if they were probability samples?

No. Of course not.

Not All "Samples" Are Alike

- Different types:

Not All "Samples" Are Alike

- Different types:
  - Passive/opt-in/"river sampling"

Not All “Samples” Are Alike

- Different types:
    - Passive/opt-in/“river sampling”
    - Sample of convenience
        - Snowball sample
        - Students
        - Crowdsourcing

Not All "Samples" Are Alike

- Different types:
  - Passive/opt-in/"river sampling"
  - Sample of convenience
    - Snowball sample
    - Students
    - Crowdsourcing
- Differ in numerous ways
  - Cost
  - "Experience"
  - Attentiveness
  - Demographics

Costs per participant

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

Participant Experience

- A lot of growing concern about experience
- Larger literature on "panel conditioning"
  - Inconclusive evidence

Participant Experience

- A lot of growing concern about experience
- Larger literature on "panel conditioning"
    - Inconclusive evidence

Participant Experience

- A lot of growing concern about experience
- Larger literature on "panel conditioning"
    - Inconclusive evidence
- Some numbers:
    - MTurk workers are doing 100+ studies per month

Participant Experience

- A lot of growing concern about experience
- Larger literature on “panel conditioning”
  - Inconclusive evidence
- Some numbers:
  - MTurk workers are doing 100+ studies per month
  - Numbers are the same for YouGov panelists

Reweighting

- If effects are heterogeneous, it may be possible to *reweight* unrepresentative 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

Review 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 within each stratum
6. Aggregate our results

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 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

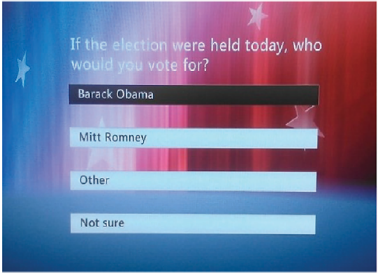
Group	Pop.	Sample	Rep.	Weight
Native-born, Female	.45	.5	Over	0.900
Native-born, Male	.45	.4	Under	1.125
Immigrant, Female	.05	.07	Over	0.714
Immigrant, Male	.05	.03	Under	1.667

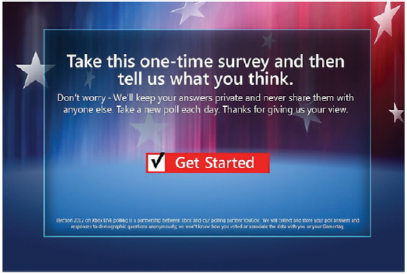
- PS weight is just $w_{ps} = N_l / n_l$

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

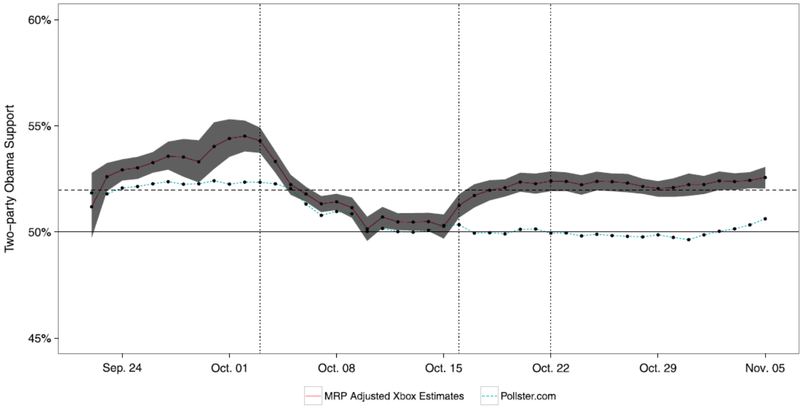
The Xbox Study



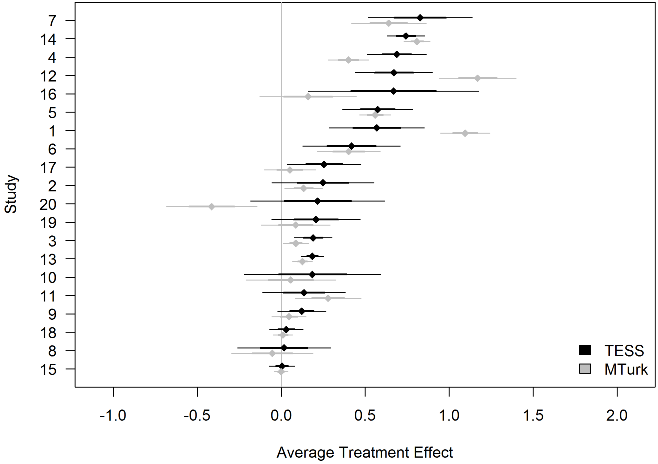


Wang et al. 2015. "Forecasting elections with non-representative polls." *International Journal of Forecasting*.

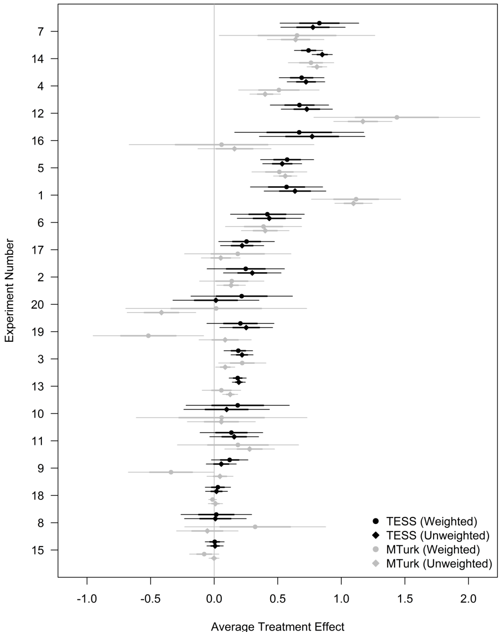
The Xbox Study



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



Mullinix et al. 2015. "The Generalizability of Survey Experiments." *Journal of Experimental Political Science*.



7
14
4
12
16
5
1
6
17
2
20
19
3
13
10
11
9
18
8
15
Experiment Number
TESS (Weighted)
TESS (Unweighted)
MTurk (Weighted)
MTurk (Unweighted)
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
Average Treatment Effect

Propensity Score Approach

1. Define a target population
2. Estimate a propensity score model
   - Pool experimental samples and target population units
   - Predict membership of all target and sample units in the experimental sample
3. Using fitted logits, divide population & sample into strata
4. Estimate stratum-specific ATE
5. Calculate weighted average of stratum-level estimates

Propensity Score Approach

Target population average treatment effect:

$$\sum_{v=1}^{5} p(v)T(v) \tag{1}$$

where $p(v)$ is the proportion of the target population in a given stratum, v, and $T(v)$ is the estimated effect from stratum v of the experimental sample

Propensity Score Approach

Effect variance:

$$\sum_{v=1}^{5} p(v)^2 V(v), \tag{2}$$

where $V(v)$ is the variance of the estimated experimental sample effect for stratum v

Propensity Score Subclassification Estimator

	Weights		Estimates			
Stratum	Nat'l	Sample	Loan	DREAM 1	DREAM 2	Rally
1	0.20	0.83	0.94 (0.08)	0.06 (0.11)	-0.22 (0.12)	0.74 (0.10)
2	0.20	0.11	0.99 (0.26)	0.22 (0.37)	-0.28 (0.36)	0.77 (0.29)
3	0.20	0.04	1.28 (0.43)	-0.61 (0.58)	-1.76 (0.54)	1.00 (0.45)
4	0.20	0.01	1.99 (0.73)	0.29 (1.12)	0.56 (0.89)	1.44 (0.79)
5	0.20	0.00				
Sample	-	-	1.04 (0.30)	-0.01 (0.44)	-0.34 (0.38)	0.79 (0.33)
Nat'l	-	-	1.14 (0.18)	0.02 (0.22)	-0.94 (0.23)	0.94 (0.19)

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 there be?
  - 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 there be?
  - What reweighting might worse bias?
- Non-coverage is a potential 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 there be?
  - What reweighting might worse bias?
- Non-coverage is a potential problem
- Not well-tested on experimental data

Questions?

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?

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." [1]

---

[1] 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[2]

[2] 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[2]
- Consequences:
    - Biased effect estimates

[2] 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[2]
- Consequences:
  - Biased effect estimates
- Mitigation:
  - Measure pretreatment
  - Avoid "pretreated" treatments or contexts
  - Study units not already treated
  - Theorize repeated effects

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

Behavioral Outcomes

- Survey experiments can rarely measure *behavior*, only *attitudes* or *behavioral intentions*
- Consequences:
  - Lack of external validity
  - Overestimates (typically) of behavioral intentions or past behavior
- Mitigation:
  - Acknowledge limitations
  - Incentivized surveys or games
  - Small behaviors

Questions?

Small Group Activity!

In groups of 4–5, consider examples from TESS, Tuesday's lecture, or your own experiences. Discuss:

- What was the researcher's question? How did they test it experimentally?
- Thinking of SUTO, in what ways is the study externally valid? In what ways is it not externally valid?

Take about 7–8 minutes.

1. External Validity of a Sample
   - Design-based
   - Model-based
2. Other Notions of External Validity
3. Other Survey Experimental Designs

Beyond One-shot Designs

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

Beyond One-shot Designs

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

I. Effect Duration

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

II. Within-Subjects Designs

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

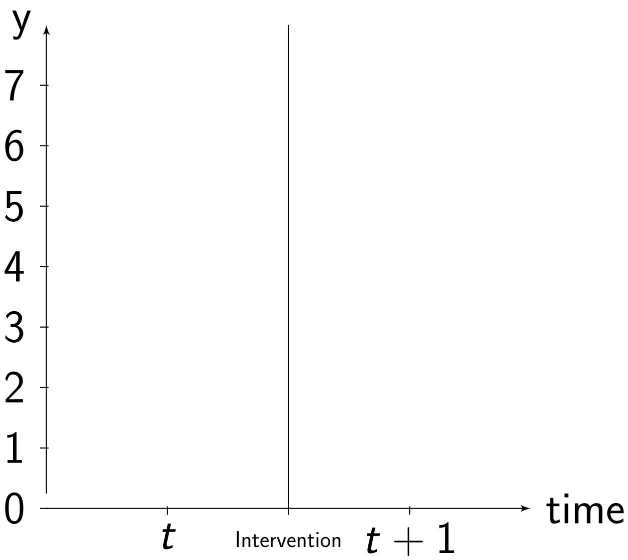
$$(\hat{Y}_{0,t+1} - \hat{Y}_{0,t}) - (\hat{Y}_{j,t+1} - \hat{Y}_{j,t})$$

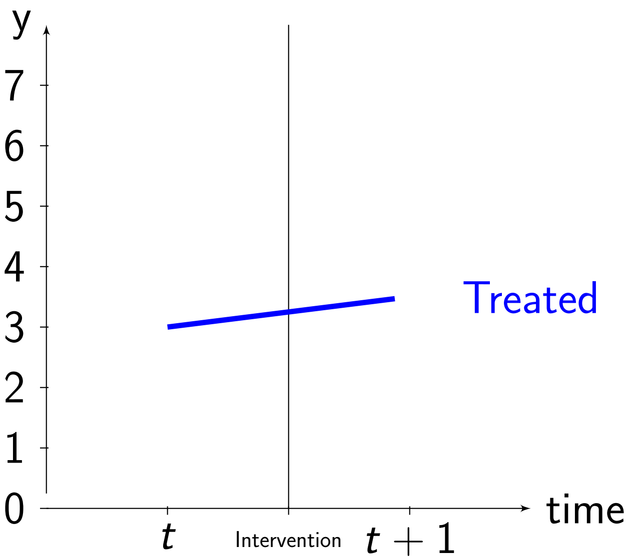
II. Within-Subjects Designs

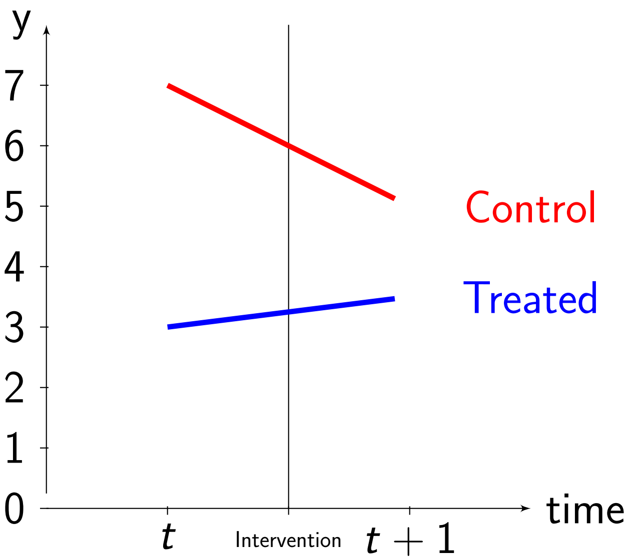
- Estimate treatment effects as a difference-in-differences
- Instead of using the post-treatment mean-difference in Y to estimate the causal effect, use the difference in pre-post differences for the two groups:

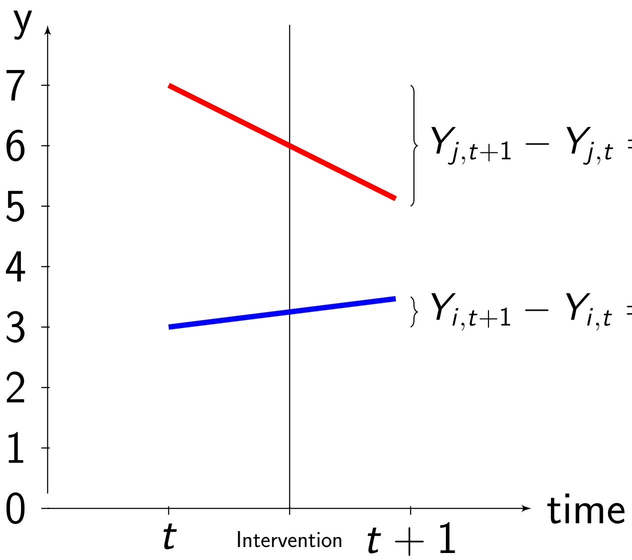
$$(\hat{Y}_{0,t+1} - \hat{Y}_{0,t}) - (\hat{Y}_{j,t+1} - \hat{Y}_{j,t})$$

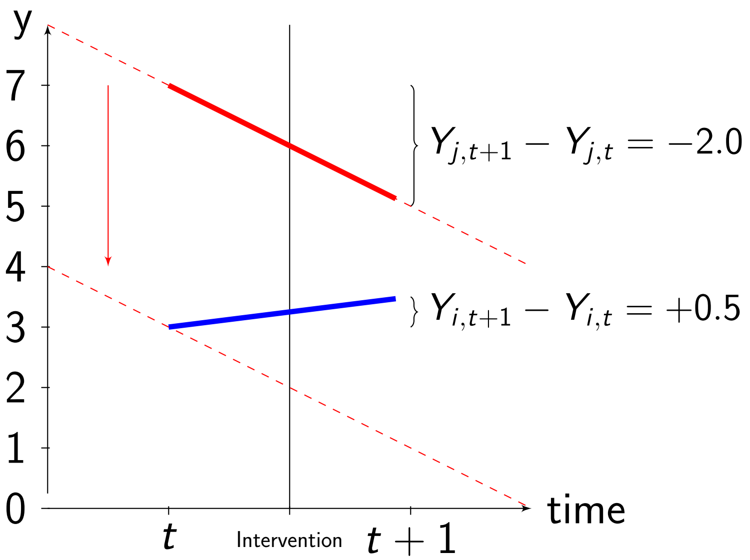
- Advantageous because variance for paired samples decreases as correlation between t_0 and t_1 observations increases

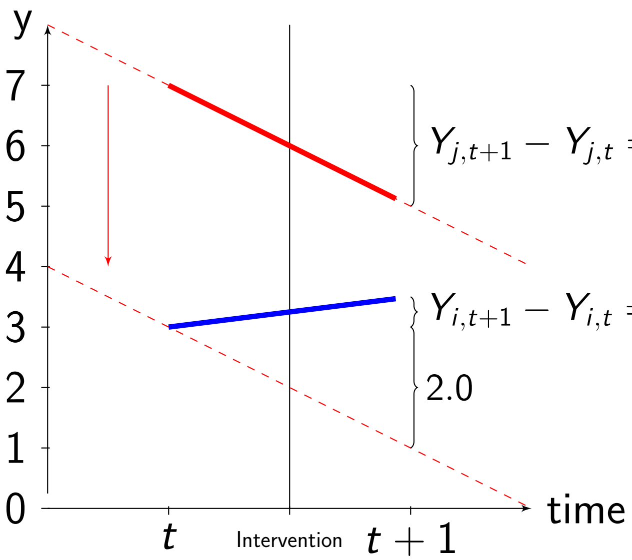

y
7
6
5
4
3
2
1
0
time
t
Intervention
$t+1$

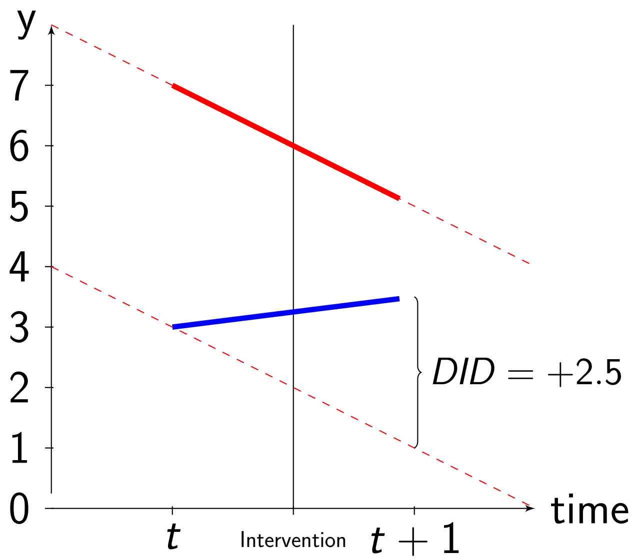

y
7
6
5
4
3
2
1
0
Treated
time
t
Intervention
$t+1$


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


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


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


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

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

Threats to Validity

As soon as time comes into play, we have to worry about threats to validity.[3]

1. History (simultaneous cause)

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

Threats to Validity

As soon as time comes into play, we have to worry about threats to validity.[3]

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

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

Threats to Validity

As soon as time comes into play, we have to worry about threats to validity.[3]

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

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

Threats to Validity

As soon as time comes into play, we have to worry about threats to validity.[3]

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

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

Threats to Validity

As soon as time comes into play, we have to worry about threats to validity.[3]

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

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

Threats to Validity

As soon as time comes into play, we have to worry about threats to validity.[3]

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

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

III. Randomized Field Treatment

- Examples:

III. Randomized Field Treatment

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

III. Randomized Field Treatment

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

III. Randomized Field Treatment

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

III. Randomized Field Treatment

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

III. Randomized Field Treatment

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

IV. Treatment Encouragement

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

IV. Treatment Encouragement

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

IV. Treatment Encouragement

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

Treatment Noncompliance

- Definition:

  "when subjects who were assigned to receive the treatment go untreated or when subjects assigned to the control group are treated" [4]

---

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

Treatment Noncompliance

- Definition:

  "when subjects who were assigned to receive the treatment go untreated or when subjects assigned to the control group are treated" [4]
- Several strategies
  - "As treated" analysis
  - "Intention to treat" analysis
  - Estimate a LATE

---

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

Analyzing Noncompliance

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

Analyzing Noncompliance

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

Local Average Treatment Effect

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

Questions?

Homework!

None! Enjoy your Wednesday!