Day 11: Design choices for (online) survey experiments

Erin Rossiter

March, 2022

Announcements

Announcements

- Emailed comments on paper drafts if you submitted
- Next week (April 5): advanced topics
  - Conjoints and list experiments (new stuff for me!)
  - Mediation
  - Anything else?
- April 12: wrap up reflection and opportunities moving forward
- April 19: final class & presentations!

Announcements

- Emailed comments on paper drafts if you submitted
- Next week (April 5): advanced topics
  - Conjoints and list experiments (new stuff for me!)
  - Mediation
  - Anything else?
- April 12: wrap up reflection and opportunities moving forward
- April 19: final class & presentations!

Announcements

- Emailed comments on paper drafts if you submitted
- Next week (April 5): advanced topics
  - Conjoints and list experiments (new stuff for me!)
  - Mediation
  - Anything else?
- April 12: wrap up reflection and opportunities moving forward
- April 19: final class & presentations!

Announcements

- Emailed comments on paper drafts if you submitted
- Next week (April 5): advanced topics
  - Conjoints and list experiments (new stuff for me!)
  - Mediation
  - Anything else?
- April 12: wrap up reflection and opportunities moving forward
- April 19: final class & presentations!

Announcements

- Emailed comments on paper drafts if you submitted
- Next week (April 5): advanced topics
  - Conjoints and list experiments (new stuff for me!)
  - Mediation
  - Anything else?
- April 12: wrap up reflection and opportunities moving forward
- April 19: final class & presentations!

Announcements

- Emailed comments on paper drafts if you submitted
- Next week (April 5): advanced topics
  - Conjoints and list experiments (new stuff for me!)
  - Mediation
  - Anything else?
- April 12: wrap up reflection and opportunities moving forward
- April 19: final class & presentations!

Announcements

- Emailed comments on paper drafts if you submitted
- Next week (April 5): advanced topics
  - Conjoints and list experiments (new stuff for me!)
  - Mediation
  - Anything else?
- April 12: wrap up reflection and opportunities moving forward
- April 19: final class & presentations!

Announcements

- Emailed comments on paper drafts if you submitted
- Next week (April 5): advanced topics
  - Conjoints and list experiments (new stuff for me!)
  - Mediation
  - Anything else?
- April 12: wrap up reflection and opportunities moving forward
- April 19: final class & presentations!

Today

Today

- online markets & survey experiments
  - goal to orient us to these discussions and literatures, but we can't do it all
  - more next week if you want
  - I'll email a zip of these papers
  - keep in mind how doable many of these papers are for methods paper inspiration
- general final paper comments
  - blocking
  - power analysis

Today

- online markets & survey experiments
  - goal to orient us to these discussions and literatures, but we can't do it all
  - more next week if you want
  - I'll email a zip of these papers
  - keep in mind how doable many of these papers are for methods paper inspiration
- general final paper comments
  - blocking
  - power analysis

Today

- online markets & survey experiments
  - goal to orient us to these discussions and literatures, but we can't do it all
  - more next week if you want
  - I'll email a zip of these papers
  - keep in mind how doable many of these papers are for methods paper inspiration
- general final paper comments
  - blocking
  - power analysis

Today

- online markets & survey experiments
  - goal to orient us to these discussions and literatures, but we can't do it all
  - more next week if you want
  - I'll email a zip of these papers
  - keep in mind how doable many of these papers are for methods paper inspiration
- general final paper comments
  - blocking
  - power analysis

Today

- online markets & survey experiments
  - goal to orient us to these discussions and literatures, but we can't do it all
  - more next week if you want
  - I'll email a zip of these papers
  - keep in mind how doable many of these papers are for methods paper inspiration
- general final paper comments
  - blocking
  - power analysis

Today

- online markets & survey experiments
  - goal to orient us to these discussions and literatures, but we can't do it all
  - more next week if you want
  - I'll email a zip of these papers
  - keep in mind how doable many of these papers are for methods paper inspiration
- general final paper comments
  - blocking
  - power analysis

Today

- online markets & survey experiments
  - goal to orient us to these discussions and literatures, but we can't do it all
  - more next week if you want
  - I'll email a zip of these papers
  - keep in mind how doable many of these papers are for methods paper inspiration
- general final paper comments
  - blocking
  - power analysis

Today

- online markets & survey experiments
  - goal to orient us to these discussions and literatures, but we can't do it all
  - more next week if you want
  - I'll email a zip of these papers
  - keep in mind how doable many of these papers are for methods paper inspiration
- general final paper comments
  - blocking
  - power analysis

Today

- online markets & survey experiments
  - goal to orient us to these discussions and literatures, but we can't do it all
  - more next week if you want
  - I'll email a zip of these papers
  - keep in mind how doable many of these papers are for methods paper inspiration
- general final paper comments
  - blocking
  - power analysis

Online markets

Convenience samples and representativeness

- Many experiments in political science use convenience samples
  - student samples
  - community samples
  - **online samples**
- Concern: what do these samples look like? Is there different demographic representation, political attitudes and behaviors, personality traits, etc?
  - Why do we care about this?

Convenience samples and representativeness

- Many experiments in political science use convenience samples
  - student samples
  - community samples
  - **online samples**
- Concern: what do these samples look like? Is there different demographic representation, political attitudes and behaviors, personality traits, etc?
  - Why do we care about this?

Convenience samples and representativeness

- Many experiments in political science use convenience samples
  - student samples
  - community samples
  - **online samples**
- Concern: what do these samples look like? Is there different demographic representation, political attitudes and behaviors, personality traits, etc?
  - Why do we care about this?

Convenience samples and representativeness

- Many experiments in political science use convenience samples
  - student samples
  - community samples
  - **online samples**
- Concern: what do these samples look like? Is there different demographic representation, political attitudes and behaviors, personality traits, etc?
  - Why do we care about this?

Convenience samples and representativeness

- Many experiments in political science use convenience samples
  - student samples
  - community samples
  - **online samples**
- Concern: what do these samples look like? Is there different demographic representation, political attitudes and behaviors, personality traits, etc?
  - Why do we care about this?

Convenience samples and representativeness

- Many experiments in political science use convenience samples
  - student samples
  - community samples
  - **online samples**
- Concern: what do these samples look like? Is there different demographic representation, political attitudes and behaviors, personality traits, etc?
  - Why do we care about this?

Convenience samples and representativeness

- Many experiments in political science use convenience samples
  - student samples
  - community samples
  - **online samples**
- Concern: what do these samples look like? Is there different demographic representation, political attitudes and behaviors, personality traits, etc?
  - Why do we care about this?

Berinsky et al (2012) – demographics

Table 2 Comparing MTurk sample with other convenience samples

		Convenience Samples			
				Adult samples (Berinsky and Kinder 2006)	
Demographics	*MTurk*	*Student samples (Kam et al. 2007)*	*Adult sample (Kam et al. 2007)*	*Experiment 1: Ann Arbor, MI*	*Experiment 2: Princeton, NJ*
Female	60.1% (2.1)	56.7% (1.3)	75.7% (4.1)	66.0%	57.1%
Age (mean years)	32.3 (0.5)	20.3 (8.2)	45.5 (.916)	42.5	45.3
Education (mean years)	14.9 (0.1)	—	5.48 (1.29)	15.1	14.9
White	83.5 (1.6)	42.5	82.2 (3.7)	81.4	72.4
Black	4.4 (0.9)			12.9	22.7
Party identification					
Democrat	40.8 (2.1)			46.1	46.5
Independent	34.1 (2.0)			20.6	17.6
Republican	16.9 (1.6)			16.3	25.8
None/other	8.2 (1.2)			17.0	10.1
N	484–551	277–1428	109	141	163

Note. Percentages except for age and education with SEs in parentheses. Adult sample from Kam et al. (2007) is for campus employee participants from their Table 1, Column 1. MTurk survey is from February/March 2010.

Berinsky et al (2012) – political variables

Table 4 Comparing MTurk sample political and psychological measures to Internet and face-to-face samples

	Internet sample		*Face-to-face samples*	
	MTurk	*ANESP*	*CPS 2008*	*ANES 2008*
Registration and turnout				
Registered	78.8% (1.7)	92.0% (0.7)	71.0% (0.2)	78.2% (1.1)
Voter turnout 2008	70.6 (2.0)	89.8 (0.5)	63.6 (0.2)	70.4 (1.1)
Party identification (mean on 7-point scale, 7 = Strong Republican)	3.48 (0.09)	3.90 (0.05)		3.70 (0.05)
Ideology (mean on 7-point scale, 7 = Strong conservative)	3.39 (0.09)	4.30 (0.05)		4.24 (0.04)
Political Interest (mean on 5-point scale, 5 = Extremely interested)	2.43 (0.04)	2.71 (0.02)		2.93 (0.03)
Political knowledge (% correct)				
Presidential succession after Vice President	70.0 (1.3)	65.2 (2.0)		
House vote percentage needed to override a veto	81.3 (1.7)	73.6 (1.3)		
Number of terms to which an individual can be elected president	96.2 (0.8)	92.8 (0.7)		
Length of a U.S. Senate term	45.0 (2.1)	37.5 (1.3)		
Number of Senators per state	85.4 (1.5)	73.2 (1.2)		
Length of a U.S. House term	50.1 (2.1)	38.9 (1.3)		
Average	71.3	63.5		
Need for cognition (mean on 0–1 scale)	.625 (0.012)	.607 (0.006)		.559 (0.009)
Need to evaluate (mean on 0–1 scale)	.628 (0.008)	.579 (0.004)		.558 (0.005)
N	506–699	1,466–2,984	92,360	1,058–2,323

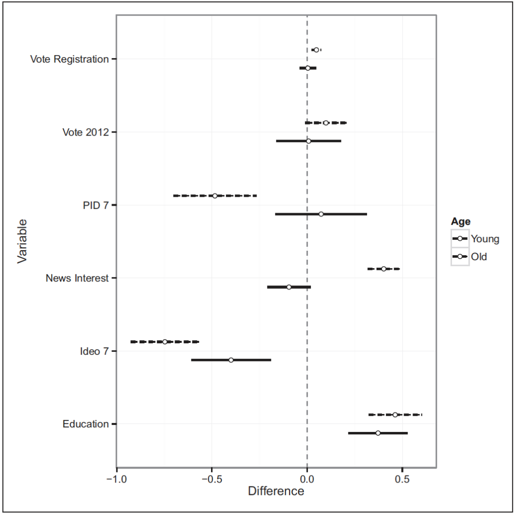
Note. Means with SEs in parentheses. CPS 2008 and ANES 2008 are weighted. Political measures are from the February/March 2010 MTurk survey ($N = 551$). Need for Cognition and Need to Evaluate are from the May 2011 MTurk survey ($N = 699$). Tests of statistical significance of differences across samples appear in the Supplementary data.

MTurk in 2012

“We demonstrate that relative to other convenience samples often used in experimental research in political science, MTurk subjects are often more representative of the general population and substantially less expensive to recruit. MTurk subjects appear to respond to experimental stimuli in a manner consistent with prior research. They are apparently also not currently an excessively overused pool, and habitual responding appears to be a minor concern. Put simply, despite possible self-selection concerns, the MTurk subject pool is no worse than convenience samples used by other researchers in political science." (Berinsky et al pg 366)

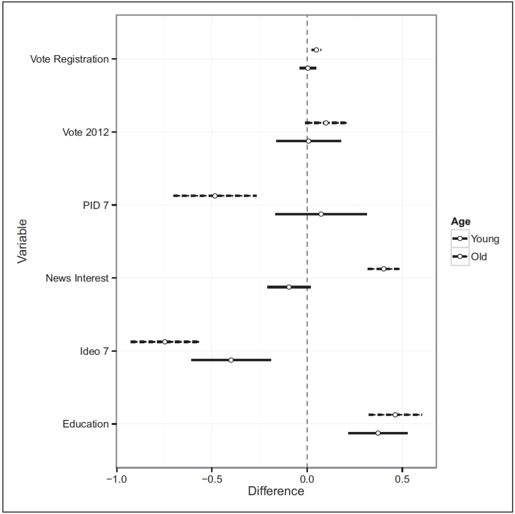
BUT - younger, more liberal, more attentive

MTurk in 2015 and external validity (Huff & Tingley 2015)



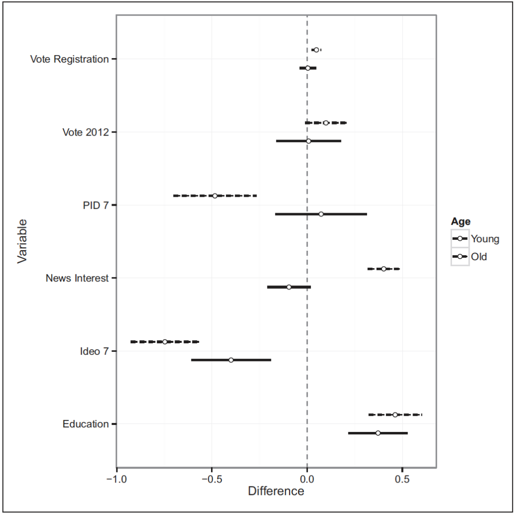
- DIM with 95% CI; Positive values indicate that MTurk > CCES
- Some manipulations and outcomes (i.e., voting tendencies) can make a better case for external validity with MTurk

MTurk in 2015 and external validity (Huff & Tingley 2015)



– DIM with 95% CI; Positive values indicate that MTurk > CCES
– Some manipulations and outcomes (i.e., voting tendencies) can make a better case for external validity with MTurk

MTurk in 2015 and external validity (Huff & Tingley 2015)



- DIM with 95% CI; Positive values indicate that MTurk > CCES
- Some manipulations and outcomes (i.e., voting tendencies) can make a better case for external validity with MTurk

External validity

- We know experiments, if meeting our assumptions, are *internally valid*
- So notice how the discussion of online samples concerns *external validity*
- **Best choice for online convenience sample will depend on study's objectives**
- Example:
  - Online experiments *about online behaviors* ought to consider older people are increasingly joining the Internet
  - they don't react the same way to it as young people (Munger et al. 2021 JEPS)
  - and, online panels like MTurk lack these kinds of people (Munger et al 2021 R&P)
    - HTE depending on age and digital literacy
    - **limits generalizability**

External validity

- We know experiments, if meeting our assumptions, are *internally valid*
- So notice how the discussion of online samples concerns *external validity*
- **Best choice for online convenience sample will depend on study's objectives**
- Example:
  - Online experiments *about online behaviors* ought to consider older people are increasingly joining the Internet
  - they don't react the same way to it as young people (Munger et al. 2021 JEPS)
  - and, online panels like MTurk lack these kinds of people (Munger et al 2021 R&P)
    - HTE depending on age and digital literacy
    - **limits generalizability**

External validity

- We know experiments, if meeting our assumptions, are *internally valid*
- So notice how the discussion of online samples concerns *external validity*
- **Best choice for online convenience sample will depend on study's objectives**
- Example:
  - Online experiments *about online behaviors* ought to consider older people are increasingly joining the Internet
  - they don't react the same way to it as young people (Munger et al. 2021 JEPS)
  - and, online panels like MTurk lack these kinds of people (Munger et al 2021 R&P)
    - HTE depending on age and digital literacy
    - **limits generalizability**

External validity

- We know experiments, if meeting our assumptions, are *internally valid*
- So notice how the discussion of online samples concerns *external validity*
- **Best choice for online convenience sample will depend on study's objectives**
- Example:
  - Online experiments *about online behaviors* ought to consider older people are increasingly joining the Internet
  - they don't react the same way to it as young people (Munger et al. 2021 JEPS)
  - and, online panels like MTurk lack these kinds of people (Munger et al 2021 R&P)
    - HTE depending on age and digital literacy
    - **limits generalizability**

External validity

- We know experiments, if meeting our assumptions, are *internally valid*
- So notice how the discussion of online samples concerns *external validity*
- **Best choice for online convenience sample will depend on study's objectives**
- Example:
  - Online experiments *about online behaviors* ought to consider older people are increasingly joining the Internet
  - they don't react the same way to it as young people (Munger et al. 2021 JEPS)
  - and, online panels like MTurk lack these kinds of people (Munger et al 2021 R&P)
    - HTE depending on age and digital literacy
    - **limits generalizability**

External validity

- We know experiments, if meeting our assumptions, are *internally valid*
- So notice how the discussion of online samples concerns *external validity*
- **Best choice for online convenience sample will depend on study's objectives**
- Example:
  - Online experiments *about online behaviors* ought to consider older people are increasingly joining the Internet
  - they don't react the same way to it as young people (Munger et al. 2021 JEPS)
  - and, online panels like MTurk lack these kinds of people (Munger et al 2021 R&P)
    - HTE depending on age and digital literacy
    - **limits generalizability**

External validity

- We know experiments, if meeting our assumptions, are *internally valid*
- So notice how the discussion of online samples concerns *external validity*
- **Best choice for online convenience sample will depend on study's objectives**
- Example:
  - Online experiments *about online behaviors* ought to consider older people are increasingly joining the Internet
  - they don't react the same way to it as young people (Munger et al. 2021 JEPS)
  - and, online panels like MTurk lack these kinds of people (Munger et al 2021 R&P)
    - HTE depending on age and digital literacy
    - **limits generalizability**

External validity

- We know experiments, if meeting our assumptions, are *internally valid*
- So notice how the discussion of online samples concerns *external validity*
- **Best choice for online convenience sample will depend on study's objectives**
- Example:
  - Online experiments *about online behaviors* ought to consider older people are increasingly joining the Internet
  - they don't react the same way to it as young people (Munger et al. 2021 JEPS)
  - and, online panels like MTurk lack these kinds of people (Munger et al 2021 R&P)
    - HTE depending on age and digital literacy
    - **limits generalizability**

External validity

- We know experiments, if meeting our assumptions, are *internally valid*
- So notice how the discussion of online samples concerns *external validity*
- **Best choice for online convenience sample will depend on study's objectives**
- Example:
  - Online experiments *about online behaviors* ought to consider older people are increasingly joining the Internet
  - they don't react the same way to it as young people (Munger et al. 2021 JEPS)
  - and, online panels like MTurk lack these kinds of people (Munger et al 2021 R&P)
    - HTE depending on age and digital literacy
    - **limits generalizability**

External validity

- We know experiments, if meeting our assumptions, are *internally valid*
- So notice how the discussion of online samples concerns *external validity*
- **Best choice for online convenience sample will depend on study's objectives**
- Example:
  - Online experiments *about online behaviors* ought to consider older people are increasingly joining the Internet
  - they don't react the same way to it as young people (Munger et al. 2021 JEPS)
  - and, online panels like MTurk lack these kinds of people (Munger et al 2021 R&P)
    - HTE depending on age and digital literacy
    - **limits generalizability**

More papers on this topic

– Krupnikov, Yanna, H. Hannah Nam, and Hillary Style. "Convenience samples in political science experiments." Advances in Experimental Political Science 165 (2021).
  - » overview in new handbook

More papers on this topic

- Krupnikov, Yanna, H. Hannah Nam, and Hillary Style. "Convenience samples in political science experiments." Advances in Experimental Political Science 165 (2021).
  - » overview in new handbook

More papers on this topic

- Krupnikov, Yanna, H. Hannah Nam, and Hillary Style. "Convenience samples in political science experiments." Advances in Experimental Political Science 165 (2021).
  - overview in new handbook

More papers on this topic

- Coppock, Alexander, Thomas J. Leeper, and Kevin J. Mullinix. "Generalizability of heterogeneous treatment effect estimates across samples." Proceedings of the National Academy of Sciences 115, no. 49 (2018): 12441-12446.
  - study 27 survey experiments and find overwhelming treatment effect *homogeneity*
  - "Our results indicate that even descriptively unrepresentative samples constructed with no design-based justification for generalizability still tend to produce useful estimates not just of the SATE but also of subgroup CATEs that generalize quite well." (mind-blown emoji!!!)
  - Coppock has more papers on this topic

More papers on this topic

- Coppock, Alexander, Thomas J. Leeper, and Kevin J. Mullinix. "Generalizability of heterogeneous treatment effect estimates across samples." Proceedings of the National Academy of Sciences 115, no. 49 (2018): 12441-12446.
  - study 27 survey experiments and find overwhelming treatment effect *homogeneity*
  - "Our results indicate that even descriptively unrepresentative samples constructed with no design-based justification for generalizability still tend to produce useful estimates not just of the SATE but also of subgroup CATEs that generalize quite well." (mind-blown emoji!!!)
  - Coppock has more papers on this topic

More papers on this topic

- Coppock, Alexander, Thomas J. Leeper, and Kevin J. Mullinix. "Generalizability of heterogeneous treatment effect estimates across samples." Proceedings of the National Academy of Sciences 115, no. 49 (2018): 12441-12446.
  - study 27 survey experiments and find overwhelming treatment effect *homogeneity*
  - "Our results indicate that even descriptively unrepresentative samples constructed with no design-based justification for generalizability still tend to produce useful estimates not just of the SATE but also of subgroup CATEs that generalize quite well." (mind-blown emoji!!!)
  - Coppock has more papers on this topic

More papers on this topic

- Coppock, Alexander, Thomas J. Leeper, and Kevin J. Mullinix. "Generalizability of heterogeneous treatment effect estimates across samples." Proceedings of the National Academy of Sciences 115, no. 49 (2018): 12441-12446.
  - study 27 survey experiments and find overwhelming treatment effect *homogeneity*
  - "Our results indicate that even descriptively unrepresentative samples constructed with no design-based justification for generalizability still tend to produce useful estimates not just of the SATE but also of subgroup CATEs that generalize quite well." (mind-blown emoji!!!)
  - Coppock has more papers on this topic

More papers on this topic

- Boas, Taylor C., Dino P. Christenson, and David M. Glick. "Recruiting large online samples in the United States and India: Facebook, mechanical turk, and qualtrics." Political Science Research and Methods 8, no. 2 (2020): 232-250.
  - need a quick, low-cost, attentive panel? Mturk
  - need to representative sample? Qualtrics
  - need to target a particular demographic group or location? recruitment via Facebook ads

More papers on this topic

- Boas, Taylor C., Dino P. Christenson, and David M. Glick. "Recruiting large online samples in the United States and India: Facebook, mechanical turk, and qualtrics." Political Science Research and Methods 8, no. 2 (2020): 232-250.
  - need a quick, low-cost, attentive panel? Mturk
  - need to representative sample? Qualtrics
  - need to target a particular demographic group or location? recruitment via Facebook ads

More papers on this topic

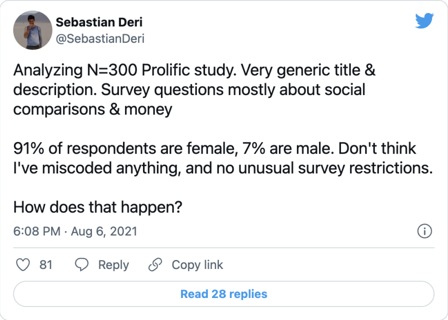
- Boas, Taylor C., Dino P. Christenson, and David M. Glick. "Recruiting large online samples in the United States and India: Facebook, mechanical turk, and qualtrics." Political Science Research and Methods 8, no. 2 (2020): 232-250.
  - need a quick, low-cost, attentive panel? Mturk
  - need to representative sample? Qualtrics
  - need to target a particular demographic group or location? recruitment via Facebook ads

More papers on this topic

- Boas, Taylor C., Dino P. Christenson, and David M. Glick. "Recruiting large online samples in the United States and India: Facebook, mechanical turk, and qualtrics." Political Science Research and Methods 8, no. 2 (2020): 232-250.
  - need a quick, low-cost, attentive panel? Mturk
  - need to representative sample? Qualtrics
  - need to target a particular demographic group or location? recruitment via Facebook ads

Caveat: this stuff changes rapidly!

Let's see an extreme example of the world of online surveys. . .

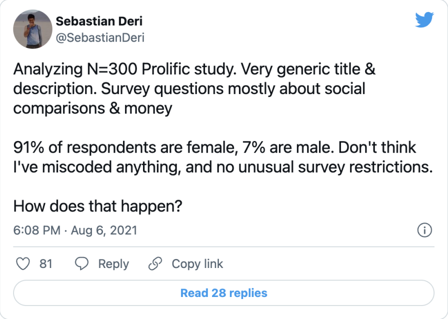


- 4.1 million views in the month after it was posted
- 10K + new users to Prolific
- 4,600 studies were disrupted (1/3 of what was on Prolific)
- no default screening tools at the time
- again, an issue with *generalizability*, not internal validity of those that were experiments

Thoughts on all this regarding your papers and projects?

Caveat: this stuff changes rapidly!

Let's see an extreme example of the world of online surveys. . .

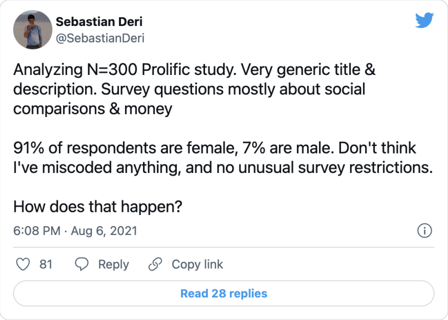


- 4.1 million views in the month after it was posted
- 10K + new users to Prolific
- 4,600 studies were disrupted (1/3 of what was on Prolific)
- no default screening tools at the time
- again, an issue with *generalizability*, not internal validity of those that were experiments

Thoughts on all this regarding your papers and projects?

Caveat: this stuff changes rapidly!

Let's see an extreme example of the world of online surveys. . .

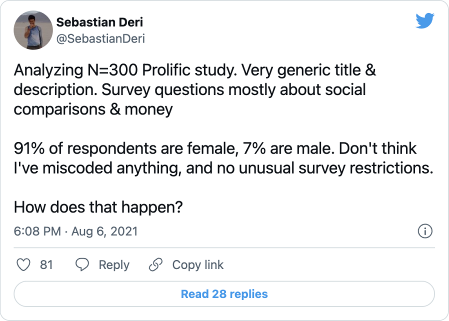


- 4.1 million views in the month after it was posted
- 10K + new users to Prolific
- 4,600 studies were disrupted (1/3 of what was on Prolific)
- no default screening tools at the time
- again, an issue with *generalizability*, not internal validity of those that were experiments

Thoughts on all this regarding your papers and projects?

Caveat: this stuff changes rapidly!

Let's see an extreme example of the world of online surveys...

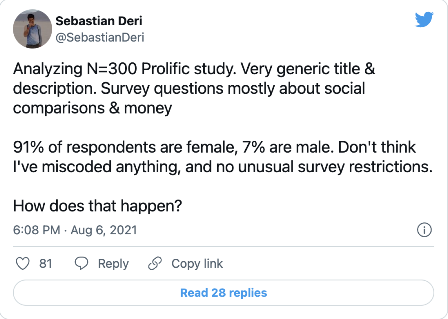


- 4.1 million views in the month after it was posted
- 10K + new users to Prolific
- 4,600 studies were disrupted (1/3 of what was on Prolific)
- no default screening tools at the time
- again, an issue with *generalizability*, not internal validity of those that were experiments

Thoughts on all this regarding your papers and projects?

Caveat: this stuff changes rapidly!

Let's see an extreme example of the world of online surveys...


Sebastian Deri
@SebastianDeri

Analyzing N=300 Prolific study. Very generic title & description. Survey questions mostly about social comparisons & money

91% of respondents are female, 7% are male. Don't think I've miscoded anything, and no unusual survey restrictions.

How does that happen?

6:08 PM · Aug 6, 2021

81 Reply Copy link

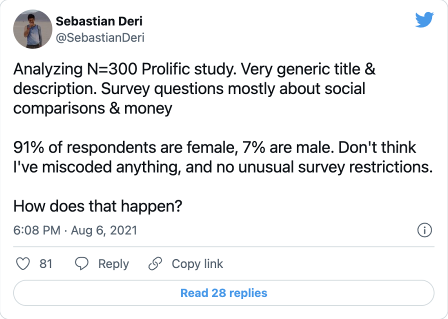
Read 28 replies

- 4.1 million views in the month after it was posted
- 10K + new users to Prolific
- 4,600 studies were disrupted (1/3 of what was on Prolific)
- no default screening tools at the time
- again, an issue with *generalizability*, not internal validity of those that were experiments

Thoughts on all this regarding your papers and projects?

Caveat: this stuff changes rapidly!

Let's see an extreme example of the world of online surveys...

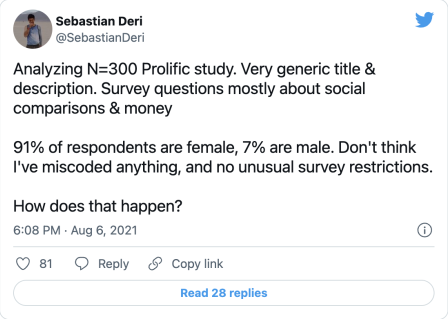


- 4.1 million views in the month after it was posted
- 10K + new users to Prolific
- 4,600 studies were disrupted (1/3 of what was on Prolific)
- no default screening tools at the time
- again, an issue with *generalizability*, not internal validity of those that were experiments

Thoughts on all this regarding your papers and projects?

Caveat: this stuff changes rapidly!

Let's see an extreme example of the world of online surveys. . .



- 4.1 million views in the month after it was posted
- 10K + new users to Prolific
- 4,600 studies were disrupted (1/3 of what was on Prolific)
- no default screening tools at the time
- again, an issue with *generalizability*, not internal validity of those that were experiments

Thoughts on all this regarding your papers and projects?

Manipulation checks

Manipulation checks

- Are participants really “treated"?
  - Harder to say with vignette experiment vs. canvassers or social interaction
- Manipulation checks = attention checks *after* treatment (in my opinion)
  - (We’ll talk about pre-treatment attention checks next)
- Why manipulation checks?
  - was the *latent* IV of interest actually affected by the stimuli?
  - Examples? From your papers?

Manipulation checks

- Are participants really “treated"?
  - Harder to say with vignette experiment vs. canvassers or social interaction
- Manipulation checks = attention checks *after* treatment (in my opinion)
  - (We’ll talk about pre-treatment attention checks next)
- Why manipulation checks?
  - was the *latent* IV of interest actually affected by the stimuli?
  - Examples? From your papers?

Manipulation checks

- Are participants really “treated"?
  - Harder to say with vignette experiment vs. canvassers or social interaction
- Manipulation checks = attention checks *after* treatment (in my opinion)
  - (We’ll talk about pre-treatment attention checks next)
- Why manipulation checks?
  - was the *latent* IV of interest actually affected by the stimuli?
  - Examples? From your papers?

Manipulation checks

- Are participants really "treated"?
  - Harder to say with vignette experiment vs. canvassers or social interaction
- Manipulation checks = attention checks *after* treatment (in my opinion)
  - (We'll talk about pre-treatment attention checks next)
- Why manipulation checks?
  - was the *latent* IV of interest actually affected by the stimuli?
  - Examples? From your papers?

Manipulation checks

- Are participants really “treated"?
  - Harder to say with vignette experiment vs. canvassers or social interaction
- Manipulation checks = attention checks *after* treatment (in my opinion)
  - (We'll talk about pre-treatment attention checks next)
- Why manipulation checks?
  - was the *latent* IV of interest actually affected by the stimuli?
  - Examples? From your papers?

Manipulation checks

- Are participants really "treated"?
  - Harder to say with vignette experiment vs. canvassers or social interaction
- Manipulation checks = attention checks *after* treatment (in my opinion)
  - (We'll talk about pre-treatment attention checks next)
- Why manipulation checks?
  - was the *latent* IV of interest actually affected by the stimuli?
  - Examples? From your papers?

Manipulation checks

- Are participants really “treated"?
  - Harder to say with vignette experiment vs. canvassers or social interaction
- Manipulation checks = attention checks *after* treatment (in my opinion)
  - (We'll talk about pre-treatment attention checks next)
- Why manipulation checks?
  - was the *latent* IV of interest actually affected by the stimuli?
  - Examples? From your papers?

Manipulation checks

- Are participants really “treated"?
  - Harder to say with vignette experiment vs. canvassers or social interaction
- Manipulation checks = attention checks *after* treatment (in my opinion)
  - (We'll talk about pre-treatment attention checks next)
- Why manipulation checks?
  - was the *latent* IV of interest actually affected by the stimuli?
  - Examples? From your papers?

Forms of manipulation checks

1. **Subjective checks**- no right or wrong answer
   - Ex: attempted to manipulate perceptions of draft reinstatement (Horowitz & Levendusky 2011)
   - "assess the likelihood that the draft will be reintroduced"
   - treatment group *should* see it as more likely than control
2. **Factual checks**- about the experimental material itself
   - Ex: attempted to manipulate whether various news stories are attributed to CNN or the Fox News Channel (Turner 2007)
   - "what network produced the stories they viewed?"
   - factually correct or incorrect answer

- Also, **instructional checks, or screeners, or attention checks**
  - anywhere in survey
  - answer is embedded in the question
  - in theory, shows who is reading closely, not just satisficing
  - could be viewed as manipulation checks, but far fetched to me
  - more on this next…

Forms of manipulation checks

1. **Subjective checks**- no right or wrong answer
   » Ex: attempted to manipulate perceptions of draft reinstatement (Horowitz & Levendusky 2011)
   » "assess the likelihood that the draft will be reintroduced"
   » treatment group *should* see it as more likely than control
2. **Factual checks**- about the experimental material itself
   » Ex: attempted to manipulate whether various news stories are attributed to CNN or the Fox News Channel (Turner 2007)
   » "what network produced the stories they viewed?"
   » factually correct or incorrect answer

– Also, **instructional checks, or screeners, or attention checks**
   » anywhere in survey
   » answer is embedded in the question
   » in theory, shows who is reading closely, not just satisficing
   » could be viewed as manipulation checks, but far fetched to me
   » more on this next. . .

Forms of manipulation checks

1. **Subjective checks**- no right or wrong answer
   - Ex: attempted to manipulate perceptions of draft reinstatement (Horowitz & Levendusky 2011)
   - "assess the likelihood that the draft will be reintroduced"
   - treatment group *should* see it as more likely than control
2. **Factual checks**- about the experimental material itself
   - Ex: attempted to manipulate whether various news stories are attributed to CNN or the Fox News Channel (Turner 2007)
   - "what network produced the stories they viewed?"
   - factually correct or incorrect answer

- Also, **instructional checks, or screeners, or attention checks**
  - anywhere in survey
  - answer is embedded in the question
  - in theory, shows who is reading closely, not just satisficing
  - could be viewed as manipulation checks, but far fetched to me
  - more on this next...

Forms of manipulation checks

1. **Subjective checks**- no right or wrong answer
   - Ex: attempted to manipulate perceptions of draft reinstatement (Horowitz & Levendusky 2011)
   - "assess the likelihood that the draft will be reintroduced"
   - treatment group *should* see it as more likely than control
2. **Factual checks**- about the experimental material itself
   - Ex: attempted to manipulate whether various news stories are attributed to CNN or the Fox News Channel (Turner 2007)
   - "what network produced the stories they viewed?"
   - factually correct or incorrect answer

- Also, **instructional checks, or screeners, or attention checks**
  - anywhere in survey
  - answer is embedded in the question
  - in theory, shows who is reading closely, not just satisficing
  - could be viewed as manipulation checks, but far fetched to me
  - more on this next...

Forms of manipulation checks

1. **Subjective checks**- no right or wrong answer
   - Ex: attempted to manipulate perceptions of draft reinstatement (Horowitz & Levendusky 2011)
   - "assess the likelihood that the draft will be reintroduced"
   - treatment group *should* see it as more likely than control
2. **Factual checks**- about the experimental material itself
   - Ex: attempted to manipulate whether various news stories are attributed to CNN or the Fox News Channel (Turner 2007)
   - "what network produced the stories they viewed?"
   - factually correct or incorrect answer

- Also, **instructional checks, or screeners, or attention checks**
  - anywhere in survey
  - answer is embedded in the question
  - in theory, shows who is reading closely, not just satisficing
  - could be viewed as manipulation checks, but far fetched to me
  - more on this next...

Forms of manipulation checks

1. **Subjective checks**- no right or wrong answer
   - Ex: attempted to manipulate perceptions of draft reinstatement (Horowitz & Levendusky 2011)
   - "assess the likelihood that the draft will be reintroduced"
   - treatment group *should* see it as more likely than control
2. **Factual checks**- about the experimental material itself
   - Ex: attempted to manipulate whether various news stories are attributed to CNN or the Fox News Channel (Turner 2007)
   - "what network produced the stories they viewed?"
   - factually correct or incorrect answer

- Also, **instructional checks, or screeners, or attention checks**
  - anywhere in survey
  - answer is embedded in the question
  - in theory, shows who is reading closely, not just satisficing
  - could be viewed as manipulation checks, but far fetched to me
  - more on this next…

Forms of manipulation checks

1. **Subjective checks**- no right or wrong answer
   - Ex: attempted to manipulate perceptions of draft reinstatement (Horowitz & Levendusky 2011)
   - "assess the likelihood that the draft will be reintroduced"
   - treatment group *should* see it as more likely than control
2. **Factual checks**- about the experimental material itself
   - Ex: attempted to manipulate whether various news stories are attributed to CNN or the Fox News Channel (Turner 2007)
   - "what network produced the stories they viewed?"
   - factually correct or incorrect answer

- Also, **instructional checks, or screeners, or attention checks**
  - anywhere in survey
  - answer is embedded in the question
  - in theory, shows who is reading closely, not just satisficing
  - could be viewed as manipulation checks, but far fetched to me
  - more on this next...

Forms of manipulation checks

1. **Subjective checks**- no right or wrong answer
   - Ex: attempted to manipulate perceptions of draft reinstatement (Horowitz & Levendusky 2011)
   - "assess the likelihood that the draft will be reintroduced"
   - treatment group *should* see it as more likely than control
2. **Factual checks**- about the experimental material itself
   - Ex: attempted to manipulate whether various news stories are attributed to CNN or the Fox News Channel (Turner 2007)
   - "what network produced the stories they viewed?"
   - factually correct or incorrect answer

- Also, **instructional checks, or screeners, or attention checks**
  - anywhere in survey
  - answer is embedded in the question
  - in theory, shows who is reading closely, not just satisficing
  - could be viewed as manipulation checks, but far fetched to me
  - more on this next...

Forms of manipulation checks

1. **Subjective checks**- no right or wrong answer
   » Ex: attempted to manipulate perceptions of draft reinstatement (Horowitz & Levendusky 2011)
   » "assess the likelihood that the draft will be reintroduced"
   » treatment group *should* see it as more likely than control
2. **Factual checks**- about the experimental material itself
   » Ex: attempted to manipulate whether various news stories are attributed to CNN or the Fox News Channel (Turner 2007)
   » "what network produced the stories they viewed?"
   » factually correct or incorrect answer

– Also, **instructional checks, or screeners, or attention checks**
   » anywhere in survey
   » answer is embedded in the question
   » in theory, shows who is reading closely, not just satisficing
   » could be viewed as manipulation checks, but far fetched to me
   » more on this next...

Forms of manipulation checks

1. **Subjective checks**- no right or wrong answer
   - Ex: attempted to manipulate perceptions of draft reinstatement (Horowitz & Levendusky 2011)
   - "assess the likelihood that the draft will be reintroduced"
   - treatment group *should* see it as more likely than control
2. **Factual checks**- about the experimental material itself
   - Ex: attempted to manipulate whether various news stories are attributed to CNN or the Fox News Channel (Turner 2007)
   - "what network produced the stories they viewed?"
   - factually correct or incorrect answer

- Also, **instructional checks, or screeners, or attention checks**
  - anywhere in survey
  - answer is embedded in the question
  - in theory, shows who is reading closely, not just satisficing
  - could be viewed as manipulation checks, but far fetched to me
  - more on this next...

Forms of manipulation checks

1. **Subjective checks**- no right or wrong answer
   - Ex: attempted to manipulate perceptions of draft reinstatement (Horowitz & Levendusky 2011)
   - “assess the likelihood that the draft will be reintroduced”
   - treatment group *should* see it as more likely than control
2. **Factual checks**- about the experimental material itself
   - Ex: attempted to manipulate whether various news stories are attributed to CNN or the Fox News Channel (Turner 2007)
   - “what network produced the stories they viewed?”
   - factually correct or incorrect answer

- Also, **instructional checks, or screeners, or attention checks**
  - anywhere in survey
  - answer is embedded in the question
  - in theory, shows who is reading closely, not just satisficing
  - could be viewed as manipulation checks, but far fetched to me
  - more on this next...

Forms of manipulation checks

1. **Subjective checks**- no right or wrong answer
   - » Ex: attempted to manipulate perceptions of draft reinstatement (Horowitz & Levendusky 2011)
   - » "assess the likelihood that the draft will be reintroduced"
   - » treatment group *should* see it as more likely than control
2. **Factual checks**- about the experimental material itself
   - » Ex: attempted to manipulate whether various news stories are attributed to CNN or the Fox News Channel (Turner 2007)
   - » "what network produced the stories they viewed?"
   - » factually correct or incorrect answer

– Also, **instructional checks, or screeners, or attention checks**
   - » anywhere in survey
   - » answer is embedded in the question
   - » in theory, shows who is reading closely, not just satisficing
   - » could be viewed as manipulation checks, but far fetched to me
   - » more on this next. . .

Forms of manipulation checks

1. **Subjective checks**- no right or wrong answer
   - Ex: attempted to manipulate perceptions of draft reinstatement (Horowitz & Levendusky 2011)
   - "assess the likelihood that the draft will be reintroduced"
   - treatment group *should* see it as more likely than control
2. **Factual checks**- about the experimental material itself
   - Ex: attempted to manipulate whether various news stories are attributed to CNN or the Fox News Channel (Turner 2007)
   - "what network produced the stories they viewed?"
   - factually correct or incorrect answer

- Also, **instructional checks, or screeners, or attention checks**
  - anywhere in survey
  - answer is embedded in the question
  - in theory, shows who is reading closely, not just satisficing
  - could be viewed as manipulation checks, but far fetched to me
  - more on this next. . .

Forms of manipulation checks

1. **Subjective checks**- no right or wrong answer
   - » Ex: attempted to manipulate perceptions of draft reinstatement (Horowitz & Levendusky 2011)
   - » "assess the likelihood that the draft will be reintroduced"
   - » treatment group *should* see it as more likely than control
2. **Factual checks**- about the experimental material itself
   - » Ex: attempted to manipulate whether various news stories are attributed to CNN or the Fox News Channel (Turner 2007)
   - » "what network produced the stories they viewed?"
   - » factually correct or incorrect answer

- Also, **instructional checks, or screeners, or attention checks**
  - » anywhere in survey
  - » answer is embedded in the question
  - » in theory, shows who is reading closely, not just satisficing
  - » could be viewed as manipulation checks, but far fetched to me
  - » more on this next...

Forms of manipulation checks

1. **Subjective checks**- no right or wrong answer
   - Ex: attempted to manipulate perceptions of draft reinstatement (Horowitz & Levendusky 2011)
   - "assess the likelihood that the draft will be reintroduced"
   - treatment group *should* see it as more likely than control
2. **Factual checks**- about the experimental material itself
   - Ex: attempted to manipulate whether various news stories are attributed to CNN or the Fox News Channel (Turner 2007)
   - "what network produced the stories they viewed?"
   - factually correct or incorrect answer

- Also, **instructional checks, or screeners, or attention checks**
  - anywhere in survey
  - answer is embedded in the question
  - in theory, shows who is reading closely, not just satisficing
  - could be viewed as manipulation checks, but far fetched to me
  - more on this next...

More on factual manipulation checks

Kane, John V., and Jason Barabas. "No harm in checking: Using factual manipulation checks to assess attentiveness in experiments." American Journal of Political Science 63, no. 1 (2019): 234-249.

- Introduce "factual manipulation check" typology
  - individual and group-level information on attention
  - measure attention to the thing we most care about
- Paper find no ordering effects on ATE estimates!
  - can place manipulation check before or after outcome measures
- Examples for your papers?

More on factual manipulation checks

Kane, John V., and Jason Barabas. “No harm in checking: Using factual manipulation checks to assess attentiveness in experiments.” American Journal of Political Science 63, no. 1 (2019): 234-249.

- Introduce “factual manipulation check” typology
  - individual and group-level information on attention
  - measure attention to the thing we most care about
- Paper find no ordering effects on ATE estimates!
  - can place manipulation check before or after outcome measures
- Examples for your papers?

More on factual manipulation checks

Kane, John V., and Jason Barabas. "No harm in checking: Using factual manipulation checks to assess attentiveness in experiments." American Journal of Political Science 63, no. 1 (2019): 234-249.

- Introduce "factual manipulation check" typology
  - individual and group-level information on attention
  - measure attention to the thing we most care about
- Paper find no ordering effects on ATE estimates!
  - can place manipulation check before or after outcome measures
- Examples for your papers?

More on factual manipulation checks

Kane, John V., and Jason Barabas. "No harm in checking: Using factual manipulation checks to assess attentiveness in experiments." American Journal of Political Science 63, no. 1 (2019): 234-249.

- Introduce "factual manipulation check" typology
  - individual and group-level information on attention
  - measure attention to the thing we most care about
- Paper find no ordering effects on ATE estimates!
  - can place manipulation check before or after outcome measures
- Examples for your papers?

More on factual manipulation checks

Kane, John V., and Jason Barabas. "No harm in checking: Using factual manipulation checks to assess attentiveness in experiments." American Journal of Political Science 63, no. 1 (2019): 234-249.

- Introduce "factual manipulation check" typology
  - individual and group-level information on attention
  - measure attention to the thing we most care about
- Paper find no ordering effects on ATE estimates!
  - can place manipulation check before or after outcome measures
- Examples for your papers?

More on factual manipulation checks

Kane, John V., and Jason Barabas. "No harm in checking: Using factual manipulation checks to assess attentiveness in experiments." American Journal of Political Science 63, no. 1 (2019): 234-249.

- Introduce "factual manipulation check" typology
  - individual and group-level information on attention
  - measure attention to the thing we most care about
- Paper find no ordering effects on ATE estimates!
  - can place manipulation check before or after outcome measures
- Examples for your papers?

More on factual manipulation checks

Kane, John V., and Jason Barabas. "No harm in checking: Using factual manipulation checks to assess attentiveness in experiments." American Journal of Political Science 63, no. 1 (2019): 234-249.

- Introduce "factual manipulation check" typology
  - individual and group-level information on attention
  - measure attention to the thing we most care about
- Paper find no ordering effects on ATE estimates!
  - can place manipulation check before or after outcome measures
- Examples for your papers?

Should we drop those who fail the check?

Almost always the answer is **no**!

- biases treatment effect estimates
- undermines causal identification
- basically a story about conditioning on posttreatment variables and differential attrition

Instead

- estimate ITT (effect of being *assigned* to treatment)
- ask pre-treatment attention check (discussed later)

Aronow, Peter M., Jonathon Baron, and Lauren Pinson. "A note on dropping experimental subjects who fail a manipulation check." Political Analysis 27, no. 4 (2019): 572-589.

Should we drop those who fail the check?

Almost always the answer is **no**!

- biases treatment effect estimates
- undermines causal identification
- basically a story about conditioning on posttreatment variables and differential attrition

Instead

- estimate ITT (effect of being *assigned* to treatment)
- ask pre-treatment attention check (discussed later)

Aronow, Peter M., Jonathon Baron, and Lauren Pinson. "A note on dropping experimental subjects who fail a manipulation check." Political Analysis 27, no. 4 (2019): 572-589.

Should we drop those who fail the check?

Almost always the answer is **no**!

- biases treatment effect estimates
- undermines causal identification
- basically a story about conditioning on posttreatment variables and differential attrition

Instead

- estimate ITT (effect of being *assigned* to treatment)
- ask pre-treatment attention check (discussed later)

Aronow, Peter M., Jonathon Baron, and Lauren Pinson. "A note on dropping experimental subjects who fail a manipulation check." Political Analysis 27, no. 4 (2019): 572-589.

Should we drop those who fail the check?

Almost always the answer is **no**!

- biases treatment effect estimates
- undermines causal identification
- basically a story about conditioning on posttreatment variables and differential attrition

Instead

- estimate ITT (effect of being *assigned* to treatment)
- ask pre-treatment attention check (discussed later)

Aronow, Peter M., Jonathon Baron, and Lauren Pinson. "A note on dropping experimental subjects who fail a manipulation check." Political Analysis 27, no. 4 (2019): 572-589.

Should we drop those who fail the check?

Almost always the answer is **no**!

- biases treatment effect estimates
- undermines causal identification
- basically a story about conditioning on posttreatment variables and differential attrition

Instead

- estimate ITT (effect of being *assigned* to treatment)
- ask pre-treatment attention check (discussed later)

Aronow, Peter M., Jonathon Baron, and Lauren Pinson. "A note on dropping experimental subjects who fail a manipulation check." Political Analysis 27, no. 4 (2019): 572-589.

Should we drop those who fail the check?

Almost always the answer is **no**!

- biases treatment effect estimates
- undermines causal identification
- basically a story about conditioning on posttreatment variables and differential attrition

Instead

- estimate ITT (effect of being *assigned* to treatment)
- ask pre-treatment attention check (discussed later)

Aronow, Peter M., Jonathon Baron, and Lauren Pinson. "A note on dropping experimental subjects who fail a manipulation check." Political Analysis 27, no. 4 (2019): 572-589.

Should we drop those who fail the check?

Almost always the answer is **no**!

- biases treatment effect estimates
- undermines causal identification
- basically a story about conditioning on posttreatment variables and differential attrition

Instead

- estimate ITT (effect of being *assigned* to treatment)
- ask pre-treatment attention check (discussed later)

Aronow, Peter M., Jonathon Baron, and Lauren Pinson. "A note on dropping experimental subjects who fail a manipulation check." Political Analysis 27, no. 4 (2019): 572-589.

Post-treatment bias

Post-treatment bias

Start with an example:/pause

- IV: civic education program
- DV: voter turnout
- Control units serve as counterfactual for treated units due to random assignment
- "Conditioning on posttreatment variables eliminates the advantages of randomization" (Mongtomery et al 2018) (!!)
  - Consider using only units who passed manipulation check
  - Subsetting down to only units who passed some civic quiz (i.e., drop units who failed)
  - The problem: now the groups aren't counterfactuals for each other!
    - those who pass in control group $\neq$ those who pass in treatment group
    - those who pass in control are people *more* knowledgeable about politics because maybe the civic education program for treated worked!
    - now pre-treatment characteristsics look different amongst groups i.e., randomization is ruined!

Post-treatment bias

Start with an example:/pause

- IV: civic education program
- DV: voter turnout
- Control units serve as counterfactual for treated units due to random assignment
- "Conditioning on posttreatment variables eliminates the advantages of randomization" (Mongtomery et al 2018) (!!)
  - Consider using only units who passed manipulation check
  - Subsetting down to only units who passed some civic quiz (i.e., drop units who failed)
  - The problem: now the groups aren't counterfactuals for each other!
    - those who pass in control group $\neq$ those who pass in treatment group
    - those who pass in control are people *more* knowledgeable about politics because maybe the civic education program for treated worked!
    - now pre-treatment characteristsics look different amongst groups i.e., randomization is ruined!

Post-treatment bias

Start with an example:/pause

- IV: civic education program
- DV: voter turnout
- Control units serve as counterfactual for treated units due to random assignment
- "Conditioning on posttreatment variables eliminates the advantages of randomization" (Mongtomery et al 2018) (!!)
  - Consider using only units who passed manipulation check
  - Subsetting down to only units who passed some civic quiz (i.e., drop units who failed)
  - The problem: now the groups aren't counterfactuals for each other!
    - those who pass in control group $\neq$ those who pass in treatment group
    - those who pass in control are people *more* knowledgeable about politics because maybe the civic education program for treated worked!
    - now pre-treatment characteristsics look different amongst groups i.e., randomization is ruined!

Post-treatment bias

Start with an example:/pause

- IV: civic education program
- DV: voter turnout
- Control units serve as counterfactual for treated units due to random assignment
- "Conditioning on posttreatment variables eliminates the advantages of randomization" (Mongtomery et al 2018) (!!)
  - Consider using only units who passed manipulation check
  - Subsetting down to only units who passed some civic quiz (i.e., drop units who failed)
  - The problem: now the groups aren't counterfactuals for each other!
    - those who pass in control group $\neq$ those who pass in treatment group
    - those who pass in control are people *more* knowledgeable about politics because maybe the civic education program for treated worked!
    - now pre-treatment characteristsics look different amongst groups i.e., randomization is ruined!

Post-treatment bias

Start with an example:/pause

- IV: civic education program
- DV: voter turnout
- Control units serve as counterfactual for treated units due to random assignment
- "Conditioning on posttreatment variables eliminates the advantages of randomization" (Mongtomery et al 2018) (!!)
  - Consider using only units who passed manipulation check
  - Subsetting down to only units who passed some civic quiz (i.e., drop units who failed)
  - The problem: now the groups aren't counterfactuals for each other!
    - those who pass in control group $\neq$ those who pass in treatment group
    - those who pass in control are people *more* knowledgeable about politics because maybe the civic education program for treated worked!
    - now pre-treatment characteristsics look different amongst groups i.e., randomization is ruined!

Post-treatment bias

Start with an example:/pause

- IV: civic education program
- DV: voter turnout
- Control units serve as counterfactual for treated units due to random assignment
- "Conditioning on posttreatment variables eliminates the advantages of randomization" (Mongtomery et al 2018) (!!)
  - Consider using only units who passed manipulation check
  - Subsetting down to only units who passed some civic quiz (i.e., drop units who failed)
  - The problem: now the groups aren't counterfactuals for each other!
    - those who pass in control group $\neq$ those who pass in treatment group
    - those who pass in control are people *more* knowledgeable about politics because maybe the civic education program for treated worked!
    - now pre-treatment characteristsics look different amongst groups i.e., randomization is ruined!

Post-treatment bias

Start with an example:/pause

- IV: civic education program
- DV: voter turnout
- Control units serve as counterfactual for treated units due to random assignment
- "Conditioning on posttreatment variables eliminates the advantages of randomization" (Mongtomery et al 2018) (!!)
  - Consider using only units who passed manipulation check
  - Subsetting down to only units who passed some civic quiz (i.e., drop units who failed)
  - The problem: now the groups aren't counterfactuals for each other!
    - those who pass in control group $\neq$ those who pass in treatment group
    - those who pass in control are people *more* knowledgeable about politics because maybe the civic education program for treated worked!
    - now pre-treatment characteristsics look different amongst groups i.e., randomization is ruined!

Post-treatment bias

Start with an example:/pause

- IV: civic education program
- DV: voter turnout
- Control units serve as counterfactual for treated units due to random assignment
- "Conditioning on posttreatment variables eliminates the advantages of randomization" (Mongtomery et al 2018) (!!)
  - Consider using only units who passed manipulation check
  - Subsetting down to only units who passed some civic quiz (i.e., drop units who failed)
  - The problem: now the groups aren't counterfactuals for each other!
    - those who pass in control group $\neq$ those who pass in treatment group
    - those who pass in control are people *more* knowledgeable about politics because maybe the civic education program for treated worked!
    - now pre-treatment characteristsics look different amongst groups i.e., randomization is ruined!

Post-treatment bias

Start with an example:/pause

- IV: civic education program
- DV: voter turnout
- Control units serve as counterfactual for treated units due to random assignment
- "Conditioning on posttreatment variables eliminates the advantages of randomization" (Mongtomery et al 2018) (!!)
  - Consider using only units who passed manipulation check
  - Subsetting down to only units who passed some civic quiz (i.e., drop units who failed)
  - The problem: now the groups aren't counterfactuals for each other!
    - those who pass in control group $\neq$ those who pass in treatment group
    - those who pass in control are people *more* knowledgeable about politics because maybe the civic education program for treated worked!
    - now pre-treatment characteristsics look different amongst groups i.e., randomization is ruined!

Post-treatment bias

Start with an example:/pause

- IV: civic education program
- DV: voter turnout
- Control units serve as counterfactual for treated units due to random assignment
- "Conditioning on posttreatment variables eliminates the advantages of randomization" (Mongtomery et al 2018) (!!)
  - Consider using only units who passed manipulation check
  - Subsetting down to only units who passed some civic quiz (i.e., drop units who failed)
  - The problem: now the groups aren't counterfactuals for each other!
    - those who pass in control group $\neq$ those who pass in treatment group
    - those who pass in control are people *more* knowledgeable about politics because maybe the civic education program for treated worked!
    - now pre-treatment characteristsics look different amongst groups i.e., randomization is ruined!

Post treatment bias

- **This example generalizes!**
  - Dropping units that failed a post-treatment manipulation check
  - Adding post-treatment variables to the regression
  - Differential attrition

Post treatment bias

- **This example generalizes!**
  - Dropping units that failed a post-treatment manipulation check
  - Adding post-treatment variables to the regression
  - Differential attrition

Post treatment bias

- **This example generalizes!**
  - Dropping units that failed a post-treatment manipulation check
  - Adding post-treatment variables to the regression
  - Differential attrition

Post treatment bias

- **This example generalizes!**
  - Dropping units that failed a post-treatment manipulation check
  - Adding post-treatment variables to the regression
  - Differential attrition

Citation

Montgomery, Jacob M., Brendan Nyhan, and Michelle Torres. “How conditioning on posttreatment variables can ruin your experiment and what to do about it.” American Journal of Political Science 62, no. 3 (2018): 760-775.

What do do?

- pre-treatment attention checks
- estimate ITT and CACE estimands

Citation

Montgomery, Jacob M., Brendan Nyhan, and Michelle Torres. “How conditioning on posttreatment variables can ruin your experiment and what to do about it.” American Journal of Political Science 62, no. 3 (2018): 760-775.

What do do?

- pre-treatment attention checks
- estimate ITT and CACE estimands

Citation

Montgomery, Jacob M., Brendan Nyhan, and Michelle Torres. "How conditioning on posttreatment variables can ruin your experiment and what to do about it." American Journal of Political Science 62, no. 3 (2018): 760-775.

What do do?

- pre-treatment attention checks
- estimate ITT and CACE estimands

Pre-treatment attention checks

Pre-treatment attention checks

What are the ingredients?

- the important thing is it is pre-treatment!
- then, they take different forms. . .

How to implement

- pre-treatment! :)
- recommendation is to use several
- should pre-register in PAP

Pre-treatment attention checks

What are the ingredients?

- the important thing is it is pre-treatment!
- then, they take different forms. . .

How to implement

- pre-treatment! :)
- recommendation is to use several
- should pre-register in PAP

Pre-treatment attention checks

What are the ingredients?

- the important thing is it is pre-treatment!
- then, they take different forms. . .

How to implement

- pre-treatment! :)
- recommendation is to use several
- should pre-register in PAP

Pre-treatment attention checks

What are the ingredients?

- the important thing is it is pre-treatment!
- then, they take different forms...

How to implement

- pre-treatment! :)
- recommendation is to use several
- should pre-register in PAP

Pre-treatment attention checks

What are the ingredients?

- the important thing is it is pre-treatment!
- then, they take different forms...

How to implement

- pre-treatment! :)
- recommendation is to use several
- should pre-register in PAP

Pre-treatment attention checks

What are the ingredients?

- the important thing is it is pre-treatment!
- then, they take different forms...

How to implement

- pre-treatment! :)
- recommendation is to use several
- should pre-register in PAP

Pre-treatment attention checks

What are the ingredients?

- the important thing is it is pre-treatment!
- then, they take different forms...

How to implement

- pre-treatment! :)
- recommendation is to use several
- should pre-register in PAP

Examples

Berinsky, Margolis and Sances 2014

FIGURE 1 An Example of a Screener Question

When a big news story breaks people often go online to get up-to-the-minute details on what is going on. We want to know which websites people trust to get this information. We also want to know if people are paying attention to the question. To show that you've read this much, please ignore the question and select ABC News and The Drudge Report as your two answers.

When there is a big news story, which is the one news website you would visit first? (Please only choose one)

☐ New York Times website	☐ The Drudge Report	☐ The Associated Press (AP) website
☐ Huffington Post	☐ Google News	☐ Reuters website
☐ Washington Post website	☐ ABC News website	☐ National Public Radio (NPR) website
☐ CNN.com	☐ CBS News website	☐ USA Today website
☐ FoxNews.com	☐ NBC News website	☐ New York Post Online
☐ MSNBC.com	☐ Yahoo! News	☐ None of these websites

Examples

Kalla and Broockman (January - May) from Aronow et al memo

- For our research, careful attention to survey questions is critical! We thank you for your care.
  - I understand
  - I do not understand

- People are very busy these days and many do not have time to follow what goes on in the government. **We are testing whether people read questions.** To show that you've read this much, answer **both** "extremely interested" **and** "very interested."
  - Extremely interested
  - Very interested
  - Moderately interested
  - Slightly interested
  - Not interested at all

Lucid and attention checks

Attention check results from Kalla and Broockman (January - May) from Aronow et al memo

Note:

- in practice, Lucid and other panel providers will allow you to drop those who failing these early attention checks without it counting against your N
- might want to ask *more* later in the survey as well

Month	N	Percent Consenting	Percent of Respondents Consenting and Passing Both Attention Checks
January	46,728	93.8%	79.9%
February	37,859	89.9%	72.8%
March	2,552	92.7%	72.6%
April	10,160	95.4%	77.1%
May	10,490	92.3%	69.8%

Lucid and attention checks

Attention check results from Kalla and Broockman (January - May) from Aronow et al memo

Note:

- in practice, Lucid and other panel providers will allow you to drop those who failing these early attention checks without it counting against your N
- might want to ask *more* later in the survey as well

Month	N	Percent Consenting	Percent of Respondents Consenting and Passing Both Attention Checks
January	46,728	93.8%	79.9%
February	37,859	89.9%	72.8%
March	2,552	92.7%	72.6%
April	10,160	95.4%	77.1%
May	10,490	92.3%	69.8%

Lucid and attention checks

Attention check results from Kalla and Broockman (January - May) from Aronow et al memo

Note:

- in practice, Lucid and other panel providers will allow you to drop those who failing these early attention checks without it counting against your N
- might want to ask *more* later in the survey as well

Month	N	Percent Consenting	Percent of Respondents Consenting and Passing Both Attention Checks
January	46,728	93.8%	79.9%
February	37,859	89.9%	72.8%
March	2,552	92.7%	72.6%
April	10,160	95.4%	77.1%
May	10,490	92.3%	69.8%

Lucid and attention checks

Attention check results from Kalla and Broockman (January - May) from Aronow et al memo

Note:

- in practice, Lucid and other panel providers will allow you to drop those who failing these early attention checks without it counting against your N
- might want to ask *more* later in the survey as well

Month	N	Percent Consenting	Percent of Respondents Consenting and Passing Both Attention Checks
January	46,728	93.8%	79.9%
February	37,859	89.9%	72.8%
March	2,552	92.7%	72.6%
April	10,160	95.4%	77.1%
May	10,490	92.3%	69.8%

More, recent advice

Berinsky, Adam J., Michele F. Margolis, Michael W. Sances, and Christopher Warshaw. "Using screeners to measure respondent attention on self-administered surveys: Which items and how many?." Political Science Research and Methods 9, no. 2 (2021): 430-437.

Advice:

- 4 questions
  - 2 grid (4 examples)
  - 2 multiple choice (4 examples)

More, recent advice

Berinsky, Adam J., Michele F. Margolis, Michael W. Sances, and Christopher Warshaw. "Using screeners to measure respondent attention on self-administered surveys: Which items and how many?." Political Science Research and Methods 9, no. 2 (2021): 430-437.

Advice:

- 4 questions
  - 2 grid (4 examples)
  - 2 multiple choice (4 examples)

More, recent advice

Berinsky, Adam J., Michele F. Margolis, Michael W. Sances, and Christopher Warshaw. "Using screeners to measure respondent attention on self-administered surveys: Which items and how many?." Political Science Research and Methods 9, no. 2 (2021): 430-437.

Advice:

- 4 questions
  - 2 grid (4 examples)
  - 2 multiple choice (4 examples)

More, recent advice

Berinsky, Adam J., Michele F. Margolis, Michael W. Sances, and Christopher Warshaw. "Using screeners to measure respondent attention on self-administered surveys: Which items and how many?." Political Science Research and Methods 9, no. 2 (2021): 430-437.

Advice:

- 4 questions
  - 2 grid (4 examples)
  - 2 multiple choice (4 examples)

More, recent advice

Berinsky, Adam J., Michele F. Margolis, Michael W. Sances, and Christopher Warshaw. "Using screeners to measure respondent attention on self-administered surveys: Which items and how many?." Political Science Research and Methods 9, no. 2 (2021): 430-437.

Advice:

- 4 questions
  - 2 grid (4 examples)
  - 2 multiple choice (4 examples)

It's never simple

Screener passage rates correlate with political variables (Berinsky, Margolis and Sances 2014)

TABLE 5 Screener Passers Differ on Observable Characteristics

	(1) 2010	(2) 2011a	(3) 2011b	(4) 2012a	(5) 2012b
# Screeners	2	3	4	2	1
Some College	0.010	−0.004	0.059	−0.006	0.051
	(0.027)	(0.034)	(0.034)	(0.034)	(0.035)
College or Above	0.104***	0.019	0.031	0.005	0.057
	(0.027)	(0.035)	(0.034)	(0.031)	(0.035)
Age	0.993**	1.563**	1.068*	0.925*	0.728
	(0.334)	(0.522)	(0.473)	(0.465)	(0.483)
Age-Squared	−5.727	−11.209*	−7.836	−7.704	−2.889
	(3.224)	(5.074)	(4.584)	(4.843)	(4.922)
Female	0.101***		0.064*	0.056*	0.120***
	(0.020)		(0.030)	(0.026)	(0.028)
Black	−0.126**	−0.196**	0.002	−0.080	−0.164***
	(0.047)	(0.061)	(0.051)	(0.047)	(0.049)
Hispanic	−0.069	0.034	−0.064	−0.107*	−0.107
	(0.073)	(0.099)	(0.085)	(0.051)	(0.061)
Other Race	−0.035	−0.149**	−0.043	−0.080	−0.141
	(0.049)	(0.055)	(0.061)	(0.053)	(0.075)
Constant	0.220*	0.070	0.329**	0.060	0.236*
	(0.087)	(0.129)	(0.120)	(0.104)	(0.114)
RMSE	0.39	0.39	0.35	0.35	0.48
R-squared	0.07	0.06	0.04	0.03	0.05
N	1,602	802	638	738	1,220

Note: All models estimated using ordinary least squares with robust standard errors in parentheses. Having a high school degree or less is

What to do in terms of internal vs external validity

If screener passage rates correlate with political variables

- dropping those who fail hurts generalizability
- including those who fail hurts internal validity

Could consider presenting heterogeneous treatment effects by attention

Always balancing these competing concerns in light of your experiment's objectives. Thoughts for your projects?

What to do in terms of internal vs external validity

If screener passage rates correlate with political variables

- dropping those who fail hurts generalizability
- including those who fail hurts internal validity

Could consider presenting heterogeneous treatment effects by attention

Always balancing these competing concerns in light of your experiment's objectives. Thoughts for your projects?

What to do in terms of internal vs external validity

If screener passage rates correlate with political variables

- dropping those who fail hurts generalizability
- including those who fail hurts internal validity

Could consider presenting heterogeneous treatment effects by attention

Always balancing these competing concerns in light of your experiment's objectives. Thoughts for your projects?

What to do in terms of internal vs external validity

If screener passage rates correlate with political variables

- dropping those who fail hurts generalizability
- including those who fail hurts internal validity

Could consider presenting heterogeneous treatment effects by attention

Always balancing these competing concerns in light of your experiment's objectives. Thoughts for your projects?

What to do in terms of internal vs external validity

If screener passage rates correlate with political variables

- dropping those who fail hurts generalizability
- including those who fail hurts internal validity

Could consider presenting heterogeneous treatment effects by attention

Always balancing these competing concerns in light of your experiment's objectives. Thoughts for your projects?

Demand effects

Demand effects

- A form of bias
  - when participants infer the purpose of an experiment and respond in a way that helps confirm the researcher's hypothesis
- Demand effects are one form of "response bias", especially worrisome in experiments.
  - Others we won't explicitly talk about that plague survey data in general:
    - social desirability bias
    - question order bias
    - acquiescence bias

Demand effects

- A form of bias
  - when participants infer the purpose of an experiment and respond in a way that helps confirm the researcher's hypothesis
- Demand effects are one form of "response bias", especially worrisome in experiments.
  - Others we won't explicitly talk about that plague survey data in general:
    - social desirability bias
    - question order bias
    - acquiescence bias

Demand effects

- A form of bias
  - when participants infer the purpose of an experiment and respond in a way that helps confirm the researcher's hypothesis
- Demand effects are one form of "response bias", especially worrisome in experiments.
  - Others we won't explicitly talk about that plague survey data in general:
    - social desirability bias
    - question order bias
    - acquiescence bias

Demand effects

- A form of bias
  - » when participants infer the purpose of an experiment and respond in a way that helps confirm the researcher's hypothesis
- Demand effects are one form of "response bias", especially worrisome in experiments.
  - » Others we won't explicitly talk about that plague survey data in general:
    - social desirability bias
    - question order bias
    - acquiescence bias

Demand effects

- A form of bias
  - when participants infer the purpose of an experiment and respond in a way that helps confirm the researcher's hypothesis
- Demand effects are one form of "response bias", especially worrisome in experiments.
  - Others we won't explicitly talk about that plague survey data in general:
    - social desirability bias
    - question order bias
    - acquiescence bias

Demand effects

- A form of bias
  - when participants infer the purpose of an experiment and respond in a way that helps confirm the researcher's hypothesis
- Demand effects are one form of "response bias", especially worrisome in experiments.
  - Others we won't explicitly talk about that plague survey data in general:
    - social desirability bias
    - question order bias
    - acquiescence bias

Demand effects

- A form of bias
  - when participants infer the purpose of an experiment and respond in a way that helps confirm the researcher's hypothesis
- Demand effects are one form of "response bias", especially worrisome in experiments.
  - Others we won't explicitly talk about that plague survey data in general:
    - social desirability bias
    - question order bias
    - acquiescence bias

Recent evidence

Mummolo, Jonathan, and Erik Peterson. "Demand effects in survey experiments: An empirical assessment." American Political Science Review 113, no. 2 (2019): 517-529.

Assess problem of demand effects using experimental approach.

- mturk participants
- they randomly assign units to know about hypothesis or not
  - doesn't alter treatment effects!
  - even when giving a bonus to respond in line with hypothesis, they don't find demand effects!
  - "Research participants exhibit a limited ability to adjust their behavior to align with researcher expectations"

Recent evidence

Mummolo, Jonathan, and Erik Peterson. "Demand effects in survey experiments: An empirical assessment." American Political Science Review 113, no. 2 (2019): 517-529.

Assess problem of demand effects using experimental approach.

- mturk participants
- they randomly assign units to know about hypothesis or not
  - doesn't alter treatment effects!
  - even when giving a bonus to respond in line with hypothesis, they don't find demand effects!
  - "Research participants exhibit a limited ability to adjust their behavior to align with researcher expectations"

Recent evidence

Mummolo, Jonathan, and Erik Peterson. "Demand effects in survey experiments: An empirical assessment." American Political Science Review 113, no. 2 (2019): 517-529.

Assess problem of demand effects using experimental approach.

- mturk participants
- they randomly assign units to know about hypothesis or not
  - doesn't alter treatment effects!
  - even when giving a bonus to respond in line with hypothesis, they don't find demand effects!
  - "Research participants exhibit a limited ability to adjust their behavior to align with researcher expectations"

Recent evidence

Mummolo, Jonathan, and Erik Peterson. "Demand effects in survey experiments: An empirical assessment." American Political Science Review 113, no. 2 (2019): 517-529.

Assess problem of demand effects using experimental approach.

- mturk participants
- they randomly assign units to know about hypothesis or not
  - doesn't alter treatment effects!
  - even when giving a bonus to respond in line with hypothesis, they don't find demand effects!
  - "Research participants exhibit a limited ability to adjust their behavior to align with researcher expectations"

Recent evidence

Mummolo, Jonathan, and Erik Peterson. "Demand effects in survey experiments: An empirical assessment." American Political Science Review 113, no. 2 (2019): 517-529.

Assess problem of demand effects using experimental approach.

- mturk participants
- they randomly assign units to know about hypothesis or not
  - doesn't alter treatment effects!
  - even when giving a bonus to respond in line with hypothesis, they don't find demand effects!
  - "Research participants exhibit a limited ability to adjust their behavior to align with researcher expectations"

Recent evidence

Mummolo, Jonathan, and Erik Peterson. "Demand effects in survey experiments: An empirical assessment." American Political Science Review 113, no. 2 (2019): 517-529.

Assess problem of demand effects using experimental approach.

- mturk participants
- they randomly assign units to know about hypothesis or not
  - doesn't alter treatment effects!
  - even when giving a bonus to respond in line with hypothesis, they don't find demand effects!
  - "Research participants exhibit a limited ability to adjust their behavior to align with researcher expectations"

Recent evidence

Mummolo, Jonathan, and Erik Peterson. "Demand effects in survey experiments: An empirical assessment." American Political Science Review 113, no. 2 (2019): 517-529.

Assess problem of demand effects using experimental approach.

- mturk participants
- they randomly assign units to know about hypothesis or not
  - doesn't alter treatment effects!
  - even when giving a bonus to respond in line with hypothesis, they don't find demand effects!
  - "Research participants exhibit a limited ability to adjust their behavior to align with researcher expectations"

Pre-test, post-test design

Pre-test, post-test design

Or, "repeated measures" design

- fear is priming or consistency pressures will affect results
  - i.e., additional causal agents, different potential outcomes than you envision
- we know this can reduce variability in potential outcomes, reducing standard error

$$SE(\hat{ATE}) = \sqrt{\frac{1}{N-1}\left(\frac{mVar(Y_i(0))}{N-m} + \frac{(N-m)Var(Y_i(1))}{m} + 2Cov(Y_i(0), Y_i(1))\right)}$$

Pre-test, post-test design

Or, "repeated measures" design

- fear is priming or consistency pressures will affect results
  - i.e., additional causal agents, different potential outcomes than you envision
- we know this can reduce variability in potential outcomes, reducing standard error

$$SE(\hat{ATE}) = \sqrt{\frac{1}{N-1}\left(\frac{mVar(Y_i(0))}{N-m} + \frac{(N-m)Var(Y_i(1))}{m} + 2Cov(Y_i(0), Y_i(1))\right)}$$

Pre-test, post-test design

Or, “repeated measures” design

- fear is priming or consistency pressures will affect results
  - i.e., additional causal agents, different potential outcomes than you envision
- we know this can reduce variability in potential outcomes, reducing standard error

$$SE(\hat{ATE}) = \sqrt{\frac{1}{N-1}\left(\frac{mVar(Y_i(0))}{N-m} + \frac{(N-m)Var(Y_i(1))}{m} + 2Cov(Y_i(0), Y_i(1))\right)}$$

Pre-test, post-test design

Or, "repeated measures" design

- fear is priming or consistency pressures will affect results
  - i.e., additional causal agents, different potential outcomes than you envision
- we know this can reduce variability in potential outcomes, reducing standard error

$$SE(\hat{ATE}) = \sqrt{\frac{1}{N-1}\left(\frac{mVar(Y_i(0))}{N-m} + \frac{(N-m)Var(Y_i(1))}{m} + 2Cov(Y_i(0), Y_i(1))\right)}$$

Pre-test, post-test design

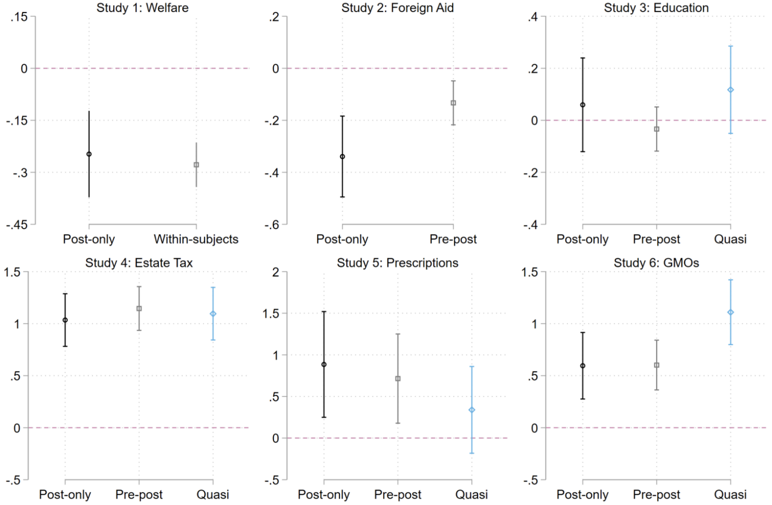
Or, "repeated measures" design

- fear is priming or consistency pressures will affect results
  - i.e., additional causal agents, different potential outcomes than you envision
- we know this can reduce variability in potential outcomes, reducing standard error

$$SE(A\hat{T}E) = \sqrt{\frac{1}{N-1}\left(\frac{mVar(Y_i(0))}{N-m} + \frac{(N-m)Var(Y_i(1))}{m} + 2Cov(Y_i(0), Y_i(1))\right)}$$

Experimental results

FIGURE 1. Treatment Effects by Experimental Design



Note: The figures display estimated average treatment effect within each design in each study. In Study 5, the displayed effect is the interaction term between the treatment and respondent partisan identity. The effects in each panel are unstandardized and plotted on the scale of the dependent variable. The bars around the estimates are 95% confidence intervals.

Multiple comparisons adjustments

Multiple comparisons adjustments

Based on EGAP 10 Things to Know by Coppock

The idea is we do *lots* of tests

- lots of treatment arms
- heterogeneous treatment effects
- different estimators (i.e., DIM with t.test or regression)
- lots of different outcomes
- lots of ways to construct outcome measures
- ....

Surely one of these will have stars!

Multiple comparisons adjustments

Based on EGAP 10 Things to Know by Coppock

The idea is we do *lots* of tests

- lots of treatment arms
- heterogeneous treatment effects
- different estimators (i.e., DIM with t.test or regression)
- lots of different outcomes
- lots of ways to construct outcome measures
- ....

Surely one of these will have stars!

Multiple comparisons adjustments

Based on EGAP 10 Things to Know by Coppock

The idea is we do *lots* of tests

- lots of treatment arms
- heterogeneous treatment effects
- different estimators (i.e., DIM with t.test or regression)
- lots of different outcomes
- lots of ways to construct outcome measures
- ....

Surely one of these will have stars!

Multiple comparisons adjustments

Based on EGAP 10 Things to Know by Coppock

The idea is we do *lots* of tests

- lots of treatment arms
- heterogeneous treatment effects
- different estimators (i.e., DIM with t.test or regression)
- lots of different outcomes
- lots of ways to construct outcome measures
- ....

Surely one of these will have stars!

Multiple comparisons adjustments

Based on EGAP 10 Things to Know by Coppock

The idea is we do *lots* of tests

- lots of treatment arms
- heterogeneous treatment effects
- different estimators (i.e., DIM with t.test or regression)
- lots of different outcomes
- lots of ways to construct outcome measures
- ....

Surely one of these will have stars!

Multiple comparisons adjustments

Based on EGAP 10 Things to Know by Coppock

The idea is we do *lots* of tests

- lots of treatment arms
- heterogeneous treatment effects
- different estimators (i.e., DIM with t.test or regression)
- lots of different outcomes
- lots of ways to construct outcome measures
- ....

Surely one of these will have stars!

Multiple comparisons adjustments

Based on EGAP 10 Things to Know by Coppock

The idea is we do *lots* of tests

- lots of treatment arms
- heterogeneous treatment effects
- different estimators (i.e., DIM with t.test or regression)
- lots of different outcomes
- lots of ways to construct outcome measures
- ....

Surely one of these will have stars!

Multiple comparisons adjustments

Based on EGAP 10 Things to Know by Coppock

The idea is we do *lots* of tests

- lots of treatment arms
- heterogeneous treatment effects
- different estimators (i.e., DIM with t.test or regression)
- lots of different outcomes
- lots of ways to construct outcome measures
- . . . .

Surely one of these will have stars!

Multiple comparisons adjustments

Based on EGAP 10 Things to Know by Coppock

The idea is we do *lots* of tests

- lots of treatment arms
- heterogeneous treatment effects
- different estimators (i.e., DIM with t.test or regression)
- lots of different outcomes
- lots of ways to construct outcome measures
- ....

Surely one of these will have stars!

Multiple comparisons adjustments

Based on EGAP 10 Things to Know by Coppock

The idea is we do *lots* of tests

- lots of treatment arms
- heterogeneous treatment effects
- different estimators (i.e., DIM with t.test or regression)
- lots of different outcomes
- lots of ways to construct outcome measures
- ....

Surely one of these will have stars!

Why the concern

Imagine testing 20 hypotheses with $\alpha = .05$. What does this mean?

- α = Pr(reject null | null is true) = false positive
- 1-α = Pr(reject null | null is false) = true positive

We take the 1/20 risk of a false positive.

If we test 20 hypotheses, one might have $p < \alpha$ *just due to chance alone*

Why the concern

Imagine testing 20 hypotheses with $\alpha = .05$. What does this mean?

- α = Pr(reject null | null is true) = false positive
- 1-α = Pr(reject null | null is false) = true positive

We take the 1/20 risk of a false positive.

If we test 20 hypotheses, one might have $p < \alpha$ *just due to chance alone*

Why the concern

Imagine testing 20 hypotheses with $\alpha = .05$. What does this mean?

- α = Pr(reject null | null is true) = false positive
- 1-α = Pr(reject null | null is false) = true positive

We take the 1/20 risk of a false positive.

If we test 20 hypotheses, one might have $p < \alpha$ *just due to chance alone*

Why the concern

Imagine testing 20 hypotheses with $\alpha = .05$. What does this mean?

- α = Pr(reject null | null is true) = false positive
- 1-α = Pr(reject null | null is false) = true positive

We take the 1/20 risk of a false positive.

If we test 20 hypotheses, one might have $p < \alpha$ *just due to chance alone*

Why the concern

Imagine testing 20 hypotheses with $\alpha = .05$. What does this mean?

- $\alpha =$ Pr(reject null | null is true) = false positive
- 1-$\alpha =$ Pr(reject null | null is false) = true positive

We take the 1/20 risk of a false positive.

If we test 20 hypotheses, one might have $p < \alpha$ *just due to chance alone*

The Bonferonni correction

Test individual hypotheses, but with significance level of $\alpha = \alpha/m$, where m is number of hypotheses. (Equivalent to adjusting pvalue*m)

```
alpha <- .05
p <- c(.0001, .02, .045, .1, .005)

bonf_p <- p.adjust(p, "bonferroni")
bonf_p
```

```
## [1] 0.0005 0.1000 0.2250 0.5000 0.0250
```

```
bonf_p < alpha
```

```
## [1]  TRUE FALSE FALSE FALSE  TRUE
```

The Bonferonni correction

Test individual hypotheses, but with significance level of $\alpha = \alpha/m$, where m is number of hypotheses. (Equivalent to adjusting pvalue*m)

```
alpha <- .05
p <- c(.0001, .02, .045, .1, .005)
```

```
bonf_p <- p.adjust(p, "bonferroni")
bonf_p
```

```
## [1] 0.0005 0.1000 0.2250 0.5000 0.0250
```

```
bonf_p < alpha
```

```
## [1]  TRUE FALSE FALSE FALSE  TRUE
```

The Bonferonni correction

Test individual hypotheses, but with significance level of $\alpha = \alpha/m$, where m is number of hypotheses. (Equivalent to adjusting pvalue*m)

```
alpha <- .05
p <- c(.0001, .02, .045, .1, .005)
```

```
bonf_p <- p.adjust(p, "bonferroni")
bonf_p
```

```
## [1] 0.0005 0.1000 0.2250 0.5000 0.0250
```

```
bonf_p < alpha
```

```
## [1]  TRUE FALSE FALSE FALSE  TRUE
```

The Bonferonni correction

Test individual hypotheses, but with significance level of $\alpha = \alpha/m$, where m is number of hypotheses. (Equivalent to adjusting pvalue*m)

```
alpha <- .05
p <- c(.0001, .02, .045, .1, .005)
```

```
bonf_p <- p.adjust(p, "bonferroni")
bonf_p
```

```
## [1] 0.0005 0.1000 0.2250 0.5000 0.0250
```

```
bonf_p < alpha
```

```
## [1]  TRUE FALSE FALSE FALSE  TRUE
```

Final papers

A review on blocking

$$SE(\hat{ATE}) = \sqrt{\frac{1}{N-1}\left(\frac{mVar(Y_i(0))}{N-m} + \frac{(N-m)Var(Y_i(1))}{m} + 2Cov(Y_i(0), Y_i(1))\right)}$$

- Recall idea behind blocking is to create blocks where units have *similar potential outcomes*
- Reduces our standard error of $\hat{ATE}$!
    - Relationship between overall uncertainty and uncertainty within each block is:
    - $SE(\hat{ATE}) = \sqrt{(}\sum_{j=1}^{J}\left(\frac{N_j}{N}\right) SE^2(\hat{ATE_j}))$
- Also beneficial for estimating CATEs because you guarantee complete random assignment even *within* the groups (no unlucky, wonky randomizations)
    - and therefore, interactions (or differences between CATEs)

A review on blocking

$$SE(\hat{ATE}) = \sqrt{\frac{1}{N-1}\left(\frac{mVar(Y_i(0))}{N-m} + \frac{(N-m)Var(Y_i(1))}{m} + 2Cov(Y_i(0), Y_i(1))\right)}$$

- Recall idea behind blocking is to create blocks where units have *similar potential outcomes*
- Reduces our standard error of $\hat{ATE}$!
  - Relationship between overall uncertainty and uncertainty within each block is:
  - $SE(\hat{ATE}) = \sqrt{(\sum_{j=1}^{J}\left(\frac{N_j}{N}\right) SE^2(\hat{ATE_j}))}$
- Also beneficial for estimating CATEs because you guarantee complete random assignment even *within* the groups (no unlucky, wonky randomizations)
  - and therefore, interactions (or differences between CATEs)

A review on blocking

$$SE(\hat{ATE}) = \sqrt{\frac{1}{N-1}\left(\frac{mVar(Y_i(0))}{N-m} + \frac{(N-m)Var(Y_i(1))}{m} + 2Cov(Y_i(0), Y_i(1))\right)}$$

- Recall idea behind blocking is to create blocks where units have *similar potential outcomes*
- Reduces our standard error of $\hat{ATE}$!
  - Relationship between overall uncertainty and uncertainty within each block is:
  - $SE(\hat{ATE}) = \sqrt{(}\sum_{j=1}^{J}\left(\frac{N_j}{N}\right) SE^2(\hat{ATE}_j))$
- Also beneficial for estimating CATEs because you guarantee complete random assignment even *within* the groups (no unlucky, wonky randomizations)
  - and therefore, interactions (or differences between CATEs)

A review on blocking

$$SE(\hat{ATE}) = \sqrt{\frac{1}{N-1}\left(\frac{mVar(Y_i(0))}{N-m} + \frac{(N-m)Var(Y_i(1))}{m} + 2Cov(Y_i(0), Y_i(1))\right)}$$

- Recall idea behind blocking is to create blocks where units have *similar potential outcomes*
- Reduces our standard error of $\hat{ATE}$!
  - Relationship between overall uncertainty and uncertainty within each block is:
  - $SE(\hat{ATE}) = \sqrt{(}\sum_{j=1}^{J}\left(\frac{N_j}{N}\right) SE^2(\hat{ATE}_j))$
- Also beneficial for estimating CATEs because you guarantee complete random assignment even *within* the groups (no unlucky, wonky randomizations)
  - and therefore, interactions (or differences between CATEs)

Simulations with DeclareDesign

Final paper should engage in "redesign" step

- probably to assess power
  - varying N and/or
  - varying effect size(s)
- if not varying N or effect size, varying something else:
  - Ex: does blocking actually help with precision?
    - compare design to complete RA and see benefits of blocking
  - Ex: what happens to power if the blocking covariate is strongly or weakly predictive of potential outcomes?
  - we can brainstorm together

Simulations with DeclareDesign

Final paper should engage in "redesign" step

- probably to assess power
  - varying N and/or
  - varying effect size(s)
- if not varying N or effect size, varying something else:
  - Ex: does blocking actually help with precision?
    - compare design to complete RA and see benefits of blocking
  - Ex: what happens to power if the blocking covariate is strongly or weakly predictive of potential outcomes?
  - we can brainstorm together

Simulations with DeclareDesign

Final paper should engage in "redesign" step

- probably to assess power
  - varying N and/or
  - varying effect size(s)
- if not varying N or effect size, varying something else:
  - Ex: does blocking actually help with precision?
    - compare design to complete RA and see benefits of blocking
  - Ex: what happens to power if the blocking covariate is strongly or weakly predictive of potential outcomes?
  - we can brainstorm together

Simulations with DeclareDesign

Final paper should engage in "redesign" step

- probably to assess power
  - varying N and/or
  - varying effect size(s)
- if not varying N or effect size, varying something else:
  - Ex: does blocking actually help with precision?
    - compare design to complete RA and see benefits of blocking
  - Ex: what happens to power if the blocking covariate is strongly or weakly predictive of potential outcomes?
  - we can brainstorm together

Simulations with DeclareDesign

Final paper should engage in "redesign" step

- probably to assess power
  - varying N and/or
  - varying effect size(s)
- if not varying N or effect size, varying something else:
  - Ex: does blocking actually help with precision?
    - compare design to complete RA and see benefits of blocking
  - Ex: what happens to power if the blocking covariate is strongly or weakly predictive of potential outcomes?
  - we can brainstorm together

Simulations with DeclareDesign

Final paper should engage in "redesign" step

- probably to assess power
  - varying N and/or
  - varying effect size(s)
- if not varying N or effect size, varying something else:
  - Ex: does blocking actually help with precision?
    - compare design to complete RA and see benefits of blocking
  - Ex: what happens to power if the blocking covariate is strongly or weakly predictive of potential outcomes?
  - we can brainstorm together

Simulations with DeclareDesign

Final paper should engage in "redesign" step

- probably to assess power
  - varying N and/or
  - varying effect size(s)
- if not varying N or effect size, varying something else:
  - Ex: does blocking actually help with precision?
    - compare design to complete RA and see benefits of blocking
  - Ex: what happens to power if the blocking covariate is strongly or weakly predictive of potential outcomes?
  - we can brainstorm together

Simulations with DeclareDesign

Final paper should engage in "redesign" step

- probably to assess power
  - varying N and/or
  - varying effect size(s)
- if not varying N or effect size, varying something else:
  - Ex: does blocking actually help with precision?
    - compare design to complete RA and see benefits of blocking
  - Ex: what happens to power if the blocking covariate is strongly or weakly predictive of potential outcomes?
  - we can brainstorm together

Simulations with DeclareDesign

Final paper should engage in "redesign" step

- probably to assess power
  - varying N and/or
  - varying effect size(s)
- if not varying N or effect size, varying something else:
  - Ex: does blocking actually help with precision?
    - compare design to complete RA and see benefits of blocking
  - Ex: what happens to power if the blocking covariate is strongly or weakly predictive of potential outcomes?
  - we can brainstorm together

Example

```
library(DeclareDesign)
```

```
## Loading required package: randomizr

## Loading required package: fabricatr

## Loading required package: estimatr
```

```
N <- 100
pop <- declare_population(N = N,
                          u = rnorm(N))
po <- declare_potential_outcomes(Y ~ .5 * Z + u)
estimand <- declare_inquiry(ATE = mean(Y_Z_1 - Y_Z_0))
assign <- declare_assignment(Z = complete_ra(N = N))
reveal <- declare_reveal()
estimator <- declare_estimator(Y ~ Z,
                               model = estimatr::difference_in_means)
design <- (pop + po + estimand
           + assign + reveal + estimator)
# this part:
design_N <- redesign(design, N = seq(100, 500, 100))
diagnosis <- diagnose_design(design_N)
```

Diagnosis varying N

```r
diagnosis$diagnosands_df[,c(1:5,10,14)]
```

```
##     design   N inquiry estimator term power mean_estimate
## 1 design_1 100     ATE estimator    Z 0.700     0.4961569
## 2 design_2 200     ATE estimator    Z 0.946     0.4998597
## 3 design_3 300     ATE estimator    Z 0.986     0.5009469
## 4 design_4 400     ATE estimator    Z 1.000     0.5046317
## 5 design_5 500     ATE estimator    Z 1.000     0.4972535
```