Generalizability of online experiments during the COVID-19 pandemic

Kyle Peyton[†]

Postdoctoral Fellow in Law and Social Science
Yale Law School
with Alex Coppock and Greg Huber

Montréal Methods Workshop
8 October 2020

[†] http://kyle-peyton.com/

Project motivation

The COVID-19 pandemic has caused significant disruptions to all aspects of life.

Fieldwork projects and in-person research largely abandoned.

Expectations from Unis are that we will all keep producing at roughly pre-COVID rates...

Many have turned toward online experimentation.

Significant disruptions raise concerns about the types of people participating in online research platforms ...

... as well as how they might respond to treatments

How much should we worry about the validity of (experimental) research conducted during this period of persistent crisis?

External validity

Concern: the (local) causal effects from experiments conducted during the pandemic may be "temporally invalid" (see Munger, 2019)

1. People behave differently. Heightened anxieties and fears over health, economic conditions, U.S. politics ...
   - Information processing biased toward threatening content (e.g. Gadarian and Albertson, 2014)
   - Decreases willingness to dissent, increases pessimism and risk-aversion (e.g. Young, 2019)
   - Increases selective exposure v. balanced information seeking (e.g. Valentino et al., 2009)
2. Extreme changes to how people spend their time (e.g. job loss), leading to changes in subject pool (e.g. Arechar and Rand, 2020)
3. Increased demand for survey respondents leading to fatigue, or the recruitment of a vastly different set of "inattentive" subjects (e.g. Aronow et al., 2020)

External validity

Question: will the (local) causal effects obtained during the pandemic generalize beyond COVID-era?

Answer strategy 1: replicate COVID-era experiments once we return to “normal times,” and compare. ~~**Answer strategy 1**: replicate COVID-era experiments once we return to “normal times,” and compare.~~

Answer strategy 2: conduct replication experiments now and compare with pre-COVID benchmarks.

Applying the UTOS framework for reasoning about generalizability (Cronbach, 1982):

1. **U**nits. Individuals that participate in online research.
2. **T**reatments. Interventions administered to units.
3. **O**utcomes. Individuals’ survey response (typically) following treatment.
4. **S**ettings. The contexts in which the above are studied.

External validity

Applying UTOS framework, we can narrow question scope:

1. Holding **T**reatments, **O**utcomes, and **S**ettings fixed, do our conclusions depend on who the **U**nits are?
   - Prior work suggests the answer is “not by much” (e.g. Mullinix et al. 2015, Coppock et al. 2019).
2. Holding **T**reatments, **O**utcomes, and **U**nits fixed, do our conclusions depend on the **S**etting (e.g. a global pandemic)?

Spoiler: we find strong evidence that pre-COVID experiments replicate (sign + significance), but at smaller magnitude.

1. We modify some **T**reatments with COVID-specific content.
   - This doesn’t seem to matter much.
2. We find evidence **U**nits have changed, and that it matters.
   - We offer suggestions for approximating pre-COVID estimates, with and without “attention checks”

Design

We conducted thirty-three replications across twelve unique studies in the **Yale Cooperative Lucid Survey (YCLS)**:

- A weekly survey of approximately 1,000 U.S. adults from Lucid, spanning 30 March to 14 July 2020
- Thirteen independent samples, each one a unique cross-section of online respondents
- Quota-sampling based on US census margins (age, gender, race/ethnicity, and region)
- “Nationally representative” on demographic marginal distributions.
- Balance on marginal distributions does not imply balance on joint distributions of demographics.
- Or balance on (joint) distribution of unobserved and unmeasured factors (major concern).

NOT A PROBABILITY SAMPLE

Selection Criteria

1. *Suitable for online survey environment*. All YCLS replications were administered via Qualtrics.
2. *Length of study*. Time constraint of 3-5 minutes.
3. *Design transparency*. Outcomes and treatments clearly described.
4. *Design complexity and effect size*. Ruled out some studies with small effect sizes, elaborate factorial designs, selective reporting, ...
5. *Theoretically important*. Most published in top journals and/or highly cited. This includes “null effects” papers.

Similar criteria to other replication projects (e.g. Many Labs)

NOT A RANDOM SAMPLE OF PRE-COVID EXPERIMENTS

Replication criteria

Criteria for declaring replication "success":

- Estimate(s) correctly signed and statistically distinguishable from zero.
- A replication "failure" occurs when estimate(s) are incorrectly signed, regardless of whether they are distinguishable from zero.
- For "null" results, replication success if estimate(s) are indistinguishable from zero, and pre-COVID benchmark.
- For studies with multiple treatment arms/outcomes, we conclude replication was successful if preponderance of evidence supports "success".
- There are edge cases that are ambiguous, and we declare these "partial replications".

Thirty-three replications conducted across twelve unique studies

Original study	Experimental design	YCLS replication	Direct replication	Replicated
Russian reporters and American news (Hyman & Sheatsley, 1950)	Two-arm	Week 3	Yes	Yes
Effect of framing on decision making (Tversky & Kaheneman, 1981)	Two-arm	Week 7	Split sample	Yes
Gain versus loss framing (Tversky & Kaheneman, 1981)	Two-arm	Weeks 1, 3, 7, 8, 13	Week 13 only	Yes
Welfare versus aid to the poor (Smith, 1987)	Two-arm	Weeks 1-9, 11-13	Yes	Yes
Gain vs. loss framing + party endorsements (Druckman, 2001)	Six-arm	Weeks 7, 8, 13	Week 13 only	Yes
Foreign aid misperceptions (Gilens, 2001)	Two-arm	Week 3	Yes	No
Perceived intentionality for side effects (Knobe, 2003)	Two-arm	Week 7	Split sample	Yes
Atomic aversion (Press, Sagan, & Valentino, 2013)	Five-arm	Weeks 5, 6, 13	Week 13 only	Partial
Attitudes towards immigrants (Hainmueller & Hopkins, 2015)	Factorial (conjoint)	Week 8	Yes	Yes
Fake news corrections (Porter, Wood, & Kirby, 2018)	Mixed factorial (2x6)	Week 4	Yes	Yes
Inequality and system justification (Trump & White, 2018)	Two-arm	Week 2	Yes	Yes
Trust in government and redistribution (Peyton, 2020)	Three-arm	Week 9	Yes	Yes

Experiment types: 1 “question order” (Russian reporters), 5 “framing” (gain v. loss, welfare v. aid to poor, etc.), 4 “information” (foreign aid, fake news, etc.), 2 “scenario evaluation” (Atomic aversion, immigration conjoint)

Results

In total, we obtained more than 90 treatment effect estimates...

- Multiple replications of same study (e.g. welfare v. aid to poor)
- Studies w/ multiple experiments (e.g. Fake News)
- Multiple outcomes + experiments (e.g. Atomic Aversion)
- Conjoint experiment yields 41 estimates alone 3

We present some individual examples, and use summary effect sizes for overall comparison between replication and pre-COVID studies:

$$\bar{\tau}_* = \frac{\sum_{i=1}^{k} w_i \hat{\tau}_i}{\sum_{i=1}^{k} w_i} \text{ with } w_i = \frac{1}{\hat{\text{se}}(\hat{\tau}_i)^2}$$

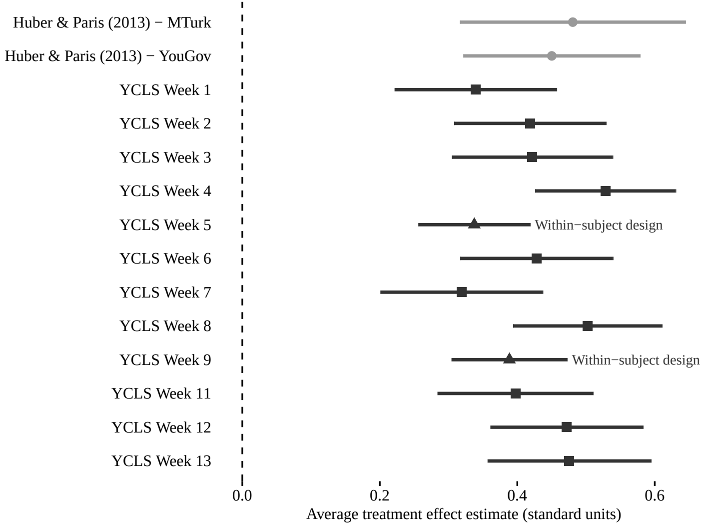
Standard error for each summary effect size: $\text{se}(\bar{\tau}_*) = \sqrt{\left(\sum_{i=1}^{k} w_i\right)^{-1}}$

Simple way to make inferences about differences between pre-COVID and replication estimates, e.g.

$$\text{se(difference)} = \sqrt{\text{se(replication)}^2 + \text{se(pre-COVID)}^2}$$

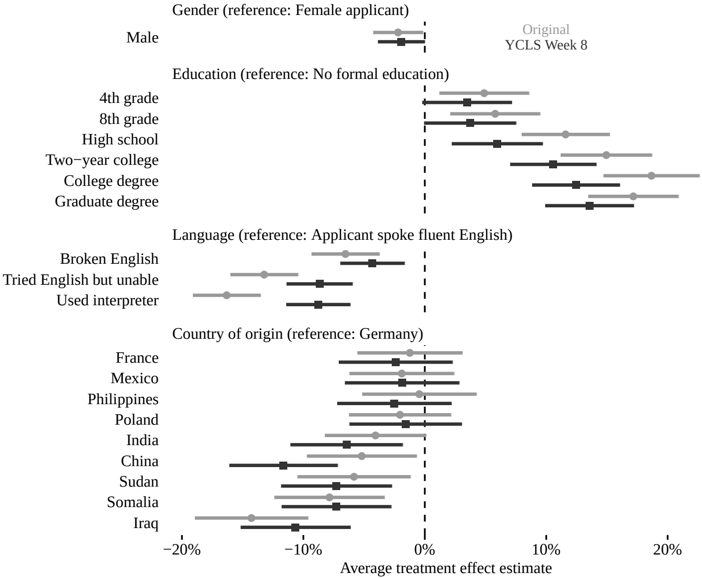
Direct replication example: government assistance

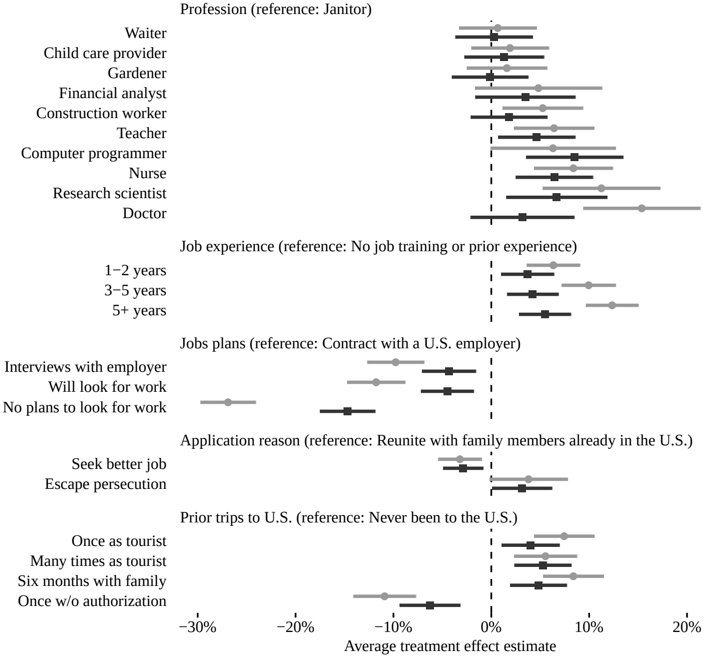
Effect of "Aid to poor" vs. "Welfare" frame on support for govt. spending



Direct replication example: conjoint experiment

Effects of immigrant attributes on support for admission to U.S.




Profession (reference: Janitor)
Waiter
Child care provider
Gardener
Financial analyst
Construction worker
Teacher
Computer programmer
Nurse
Research scientist
Doctor
Job experience (reference: No job training or prior experience)
1−2 years
3−5 years
5+ years
Jobs plans (reference: Contract with a U.S. employer)
Interviews with employer
Will look for work
No plans to look for work
Application reason (reference: Reunite with family members already in the U.S.)
Seek better job
Escape persecution
Prior trips to U.S. (reference: Never been to the U.S.)
Once as tourist
Many times as tourist
Six months with family
Once w/o authorization
−30%
−20%
−10%
0%
10%
20%
Average treatment effect estimate

COVID-specific example: modified “Asian Disease” problem

Scenario:

Imagine that the ***Mayor of a U.S. city is preparing for another outbreak of the novel coronavirus in the Spring of 2021****, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that the exact scientific estimates of the consequences of the program are as follows:*

Gain Frame:

If Program A is adopted, 200 people will be saved.

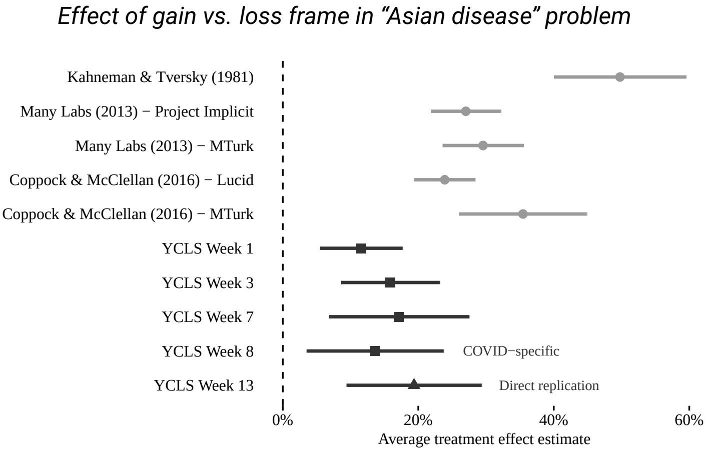
If Program B is adopted, there is a 1/3 probability that 600 people will be saved, and a 2/3 probability that no people will be saved.

Loss Frame:

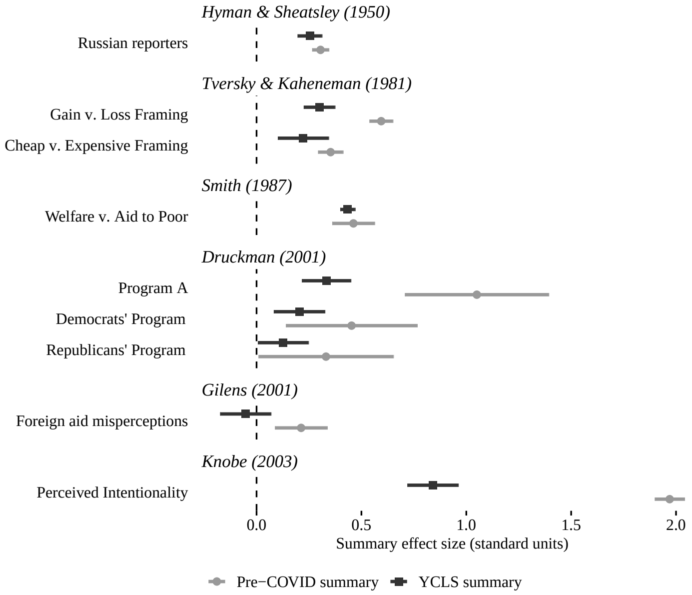
If Program A is adopted, 400 people will die.

If Program B is adopted, there is a 1/3 probability that nobody will die, and a 2/3 probability that 600 people will die.

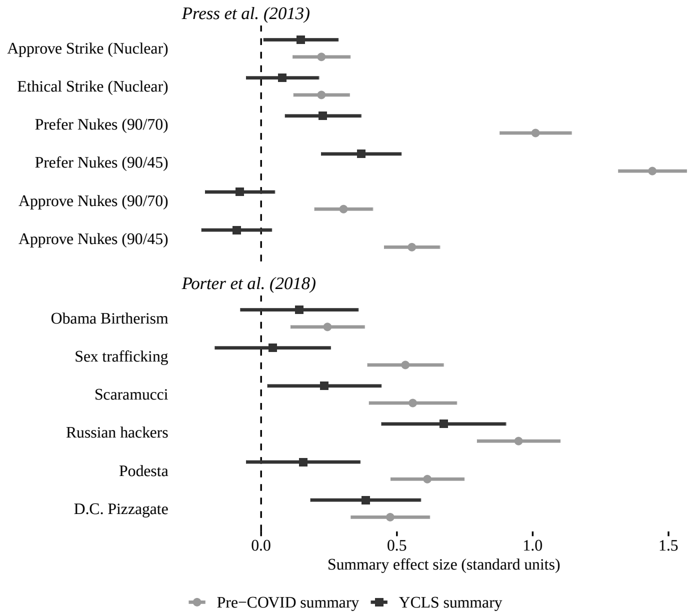
COVID-specific example: modified "Asian Disease" problem



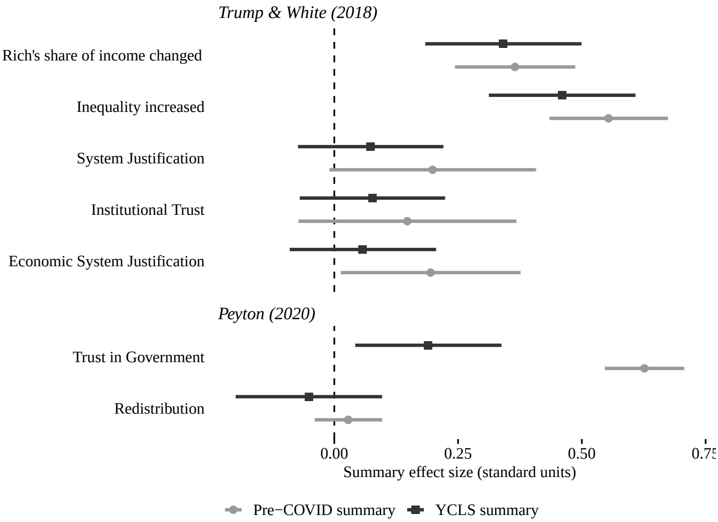
Summarizing the results (conjoint-excluded)



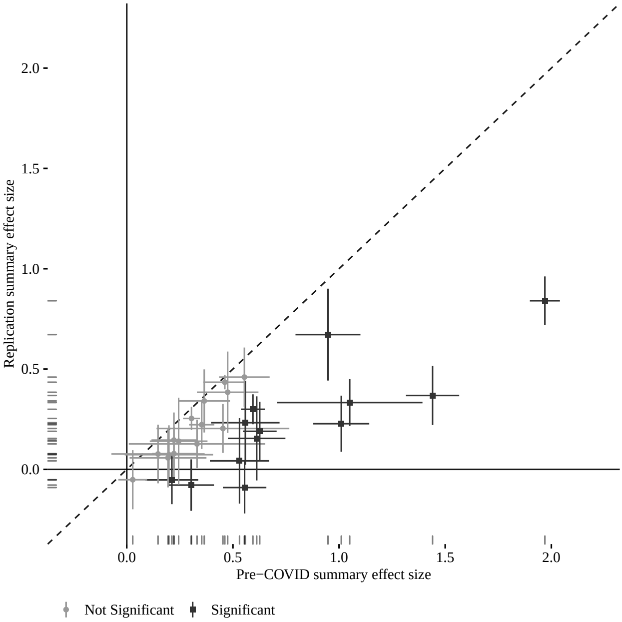
Summarizing the results (conjoint-excluded)



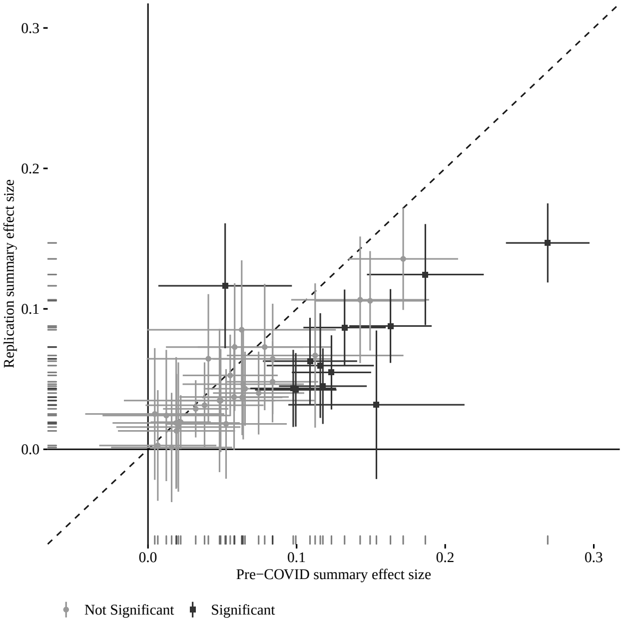
Summarizing the results (conjoint-excluded)



Comparison for 28 summary effect sizes (excludes-conjoint)



Comparison for 41 effect sizes in conjoint replication



Non-compliance framework

Type	$D_i(Z_i = 1)$	$D_i(Z_i = 0)$	$D_i(1) - D_i(0)$
Compliers	1	0	1
Never Takers	0	0	0
Always Takers	1	1	0
Defiers	0	1	-1

- **Problem:** four "types", but we cannot identify which group any particular unit belongs to.
- **Potential outcomes**: $\{Y_i(D_i(0), Z_i), Y_i(D_i(1), Z_i)\}$
- **Observed potential outcomes**:

$$Y_i = \begin{cases} Y_i(D_i = 1) & : Z_i = 1 \\ Y_i(D_i = 0) & : Z_i = 0 \\ Y_i(D_i = 0) & : Z_i = 1 \\ Y_i(D_i = 1) & : Z_i = 0 \end{cases}$$

- **One-sided noncompliance**: $D_i(0) = 1$ is ruled out.

Non-compliance framework

Type	$D_i(Z_i = 1)$	$D_i(Z_i = 0)$	$D_i(1) - D_i(0)$
Compliers	1	0	1
Never Takers	0	0	0
~~Always Takers~~	~~1~~	~~1~~	~~0~~
~~Defiers~~	~~0~~	~~1~~	~~-1~~

- **Problem:** four "types", but we cannot identify which group any particular unit belongs to.
- **Potential outcomes**: $\{Y_i(D_i(0), Z_i), Y_i(D_i(1), Z_i)\} \in \mathbb{R}$
- **Observed potential outcomes**:

$$Y_i = \begin{cases} Y_i(D_i = 1) & : Z_i = 1 \\ Y_i(D_i = 0) & : Z_i = 0 \\ Y_i(D_i = 0) & : Z_i = 1 \\ \cancel{Y_i(D_i = 1)} & : \cancel{Z_i = 0} \end{cases}$$

- **One-sided noncompliance**: $D_i(0) = 1$ is ruled out.

Non-compliance framework

Type	$D_i(Z_i = 1)$	$D_i(Z_i = 0)$	$D_i(1) - D_i(0)$
Attentive	1	0	1
Inattentive	0	0	0

- **Observed potential outcomes**:

$$Y_i = \begin{cases} Y_i(D_i = 1) & : Z_i = 1 \\ Y_i(D_i = 0) & : Z_i = 0 \\ Y_i(D_i = 0) & : Z_i = 1 \end{cases}$$

- Assume attentive reveal $Y_i(D_i = 1)$ or $Y_i(D_i = 0)$
- But inattentive reveal $Y_i(D_i = 0)$ regardless
- And $Y_i(Z) = Y_i(D_i(Z))$, regardless of type
- Replication estimates are weighted average of ATE among attentive, and 0

Causal estimands

1. Average treatment effect D_i on Y_i, for the whole sample ~~1. Average treatment effect D_i on Y_i, for the whole sample:~~

$$\mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0]$$ ~~$\mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0]$~~

2. Average effect of Z_i on Y_i, or "Intent to treat" (ITT):

$$\mathbb{E}[Y_i|Z_i = 1] - \mathbb{E}[Y_i|Z_i = 0]$$

3. "Conditional Average Treatment Effect" (CATE), among the attentive:

$$\mathbb{E}[Y_i(1) - Y_i(0)|D_i = 1] = \frac{\text{Intent to Treat Effect}}{\text{Proportion Attentive}}$$

Problem: we must estimate the denominator...

Who are the attentive?

Unlike randomized encouragement design, we cannot use Z_i to obtain unbiased estimate of π_A, or π_I...

What can we do?

1. Use **pre-treatment** attention check questions (ACQs)
   - But what if we don't have them available?
   - What if they're "too hard"? (see Berinsky et al. 2013)
2. Use others' estimates of attentiveness (e.g. Aronow et al. 2020) to re-inflate ours
3. Use auxiliary information (e.g. browser meta-data) to construct unobtrusive classifier for attentiveness

NB: these are all still estimators, no guarantees about bias or consistency.

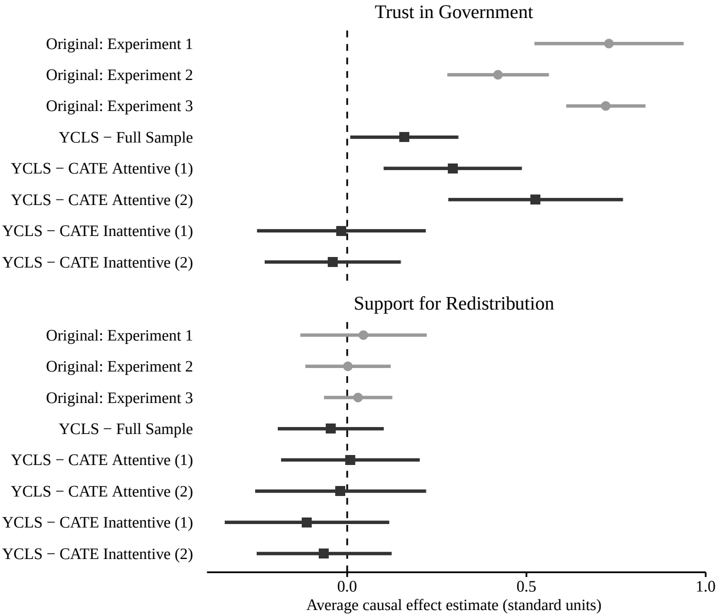
What we **should not do**: use post-treatment data to "screen out" inattentive (see Aronow et al. 2019, Montgomery et al. 2018).

Illustrations

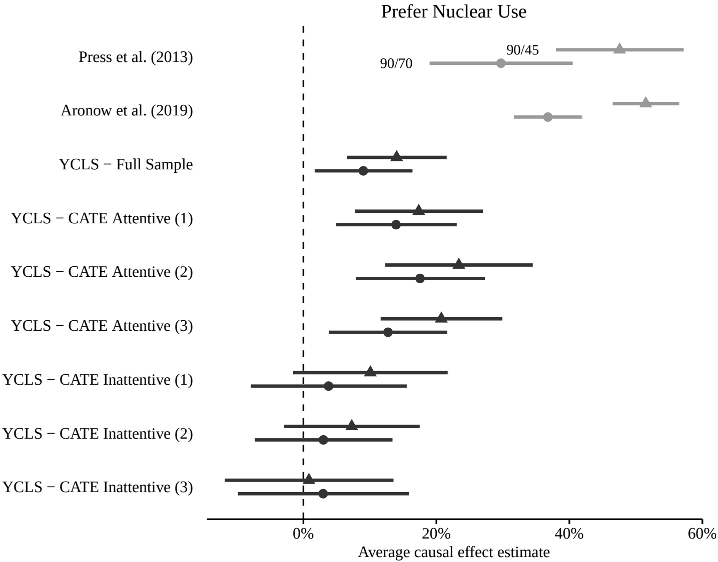
Variation in estimates for π_A by estimator across studies...

	Attentive (1)	Attentive (2)	Attentive (3)
Lucid Mar 2016	86%	60%	-
MTurk Mar 2017	99%	98%	78%
YCLS May 2020	48%	33%	22%
YCLS Jul 2020	52%	38%	60%; 45%
MTurk Jul 2020	85%	81%	-
YCLS (pooled)	56%	41%	-

Effect of corruption perceptions on trust in government and support for redistribution



Effects of relative success of "nuclear" v. "conventional" attack on support for U.S. strike on Al-Qaeda weapons lab in Syria



Summary

1. There are growing concerns about “temporal validity” of online experiments
   - The setting is certainly much different
   - Units may be different too (e.g. Arechar & Rand, 2020, Aronow et al. 2020)
2. We conducted 33 replications across 12 unique studies from Mar to Jul 2020
   - Strong evidence pre-COVID experiments replicate, but typically at reduced magnitude
   - Accumulating evidence that “inattentive” types are much more common (e.g. Lucid) than before
   - Significant increase in users coming from mobile applications (e.g. games)
3. Non-compliance framework helps us understand different estimands
   - Pre-treatment ACQs are one way to classify types. Browser meta-data is another.
   - Classifiers are not guaranteed to be unbiased or consistent

Summary

5. We believe changes in types is greater threat to generalizability than pandemic, per se
   - YCLS: $\sim$ 44% from applications v. $\sim$ 14% in 2016
   - ... $\sim$ 60% from mobile devices v. $\sim$ 40% in 2016
6. When randomly sampling potential outcomes, we know the direction of bias (e.g. attenuation) for local causal effects
   - False positive are **less** likely now than pre-COVID
   - False negative are **more** likely ...
7. Experiments must be conducted with much greater care. There is no magic “statistical fix’
   - Use ACQs, browser meta-data, etc. to estimate π_A
   - These should always be pre-treatment indicators
   - Multiple indicators? Sensitivity analysis or ensemble
   - Given variation in estimates, pre-register!

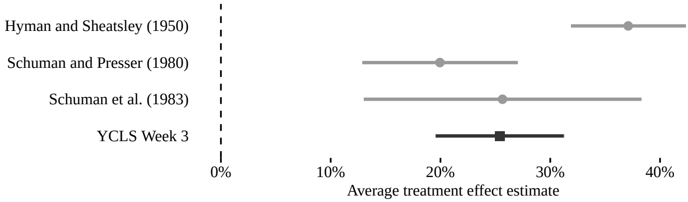
More work to be done on this. Working paper coming soon!

Thank you!

kyle.peyton@yale.edu
http://kyle-peyton.com/
@peyton_k

Original study	Experimental design	YCLS replication	Direct replication	Replicated
Russian reporters and American news (Hyman & Sheatsley, 1950)	Two-arm	Week 3	Yes	Yes
Effect of framing on decision making (Tversky & Kaheneman, 1981)	Two-arm	Week 7	Split sample	Yes
Gain versus loss framing (Tversky & Kaheneman, 1981)	Two-arm	Weeks 1, 3, 7, 8, 13	Week 13 only	Yes
Welfare versus aid to the poor (Smith, 1987)	Two-arm	Weeks 1-9, 11-13	Yes	Yes
Gain vs. loss framing + party endorsements (Druckman, 2001)	Six-arm	Weeks 7, 8, 13	Week 13 only	Yes
Foreign aid misperceptions (Gilens, 2001)	Two-arm	Week 3	Yes	No
Perceived intentionality for side effects (Knobe, 2003)	Two-arm	Week 7	Split sample	Yes
Atomic aversion (Press, Sagan, & Valentino, 2013)	Five-arm	Weeks 5, 6, 13	Week 13 only	Partial
Attitudes towards immigrants (Hainmueller & Hopkins, 2015)	Factorial (conjoint)	Week 8	Yes	Yes
Fake news corrections (Porter, Wood, & Kirby, 2018)	Mixed factorial (2x6)	Week 4	Yes	Yes
Inequality and system justification (Trump & White, 2018)	Two-arm	Week 2	Yes	Yes
Trust in government and redistribution (Peyton, 2020)	Three-arm	Week 9	Yes	Yes

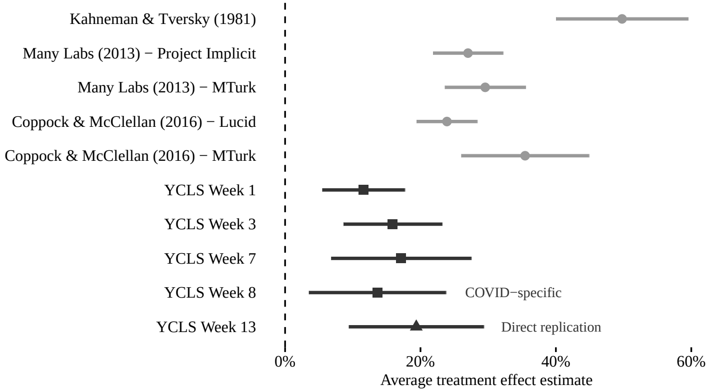
Effect of question ordering on support for Russian journalists in U.S



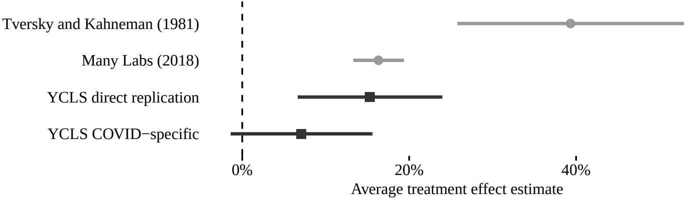
***Outcome**: "Do you think the United States should allow journalists from an authoritarian country like Russia to come in and send back the news as they see it?" [Yes = 1; No = 0]*

***Treatment**: "Do you think an authoritarian country like Russia should let American journalists come in and send back to America the news as they see it?" [First = 1; Second = 0]*

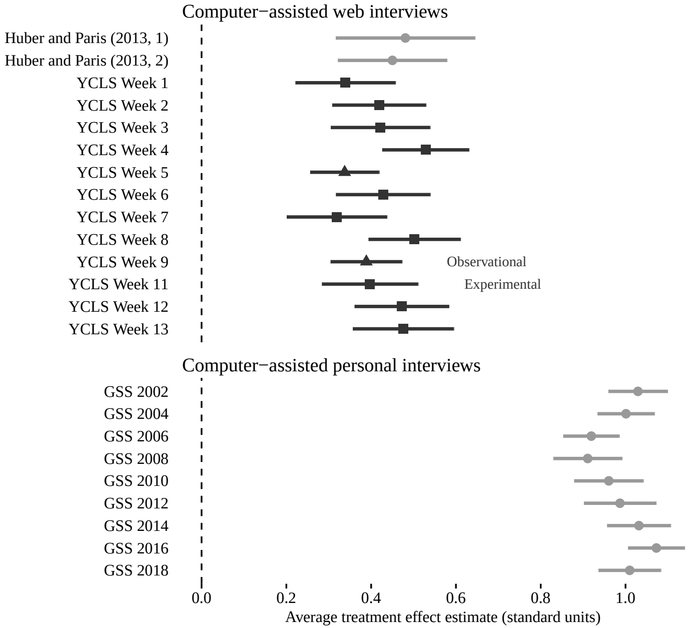
Effect of gain vs. loss frame in "Asian disease" problem



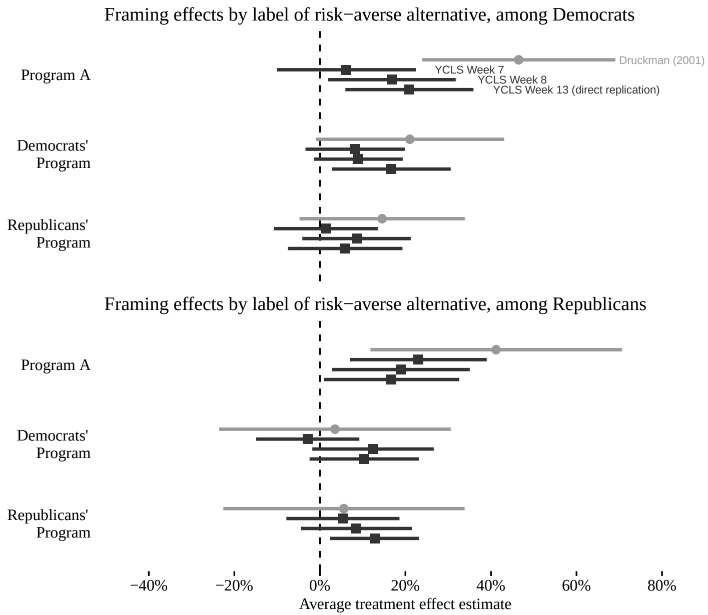
Effect of “Cheap” vs. “Expensive” frame on decision to travel



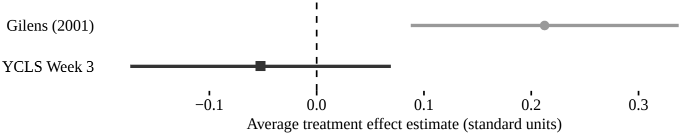
Effect of "Aid to poor" vs. "Welfare" frame on support for govt. spending



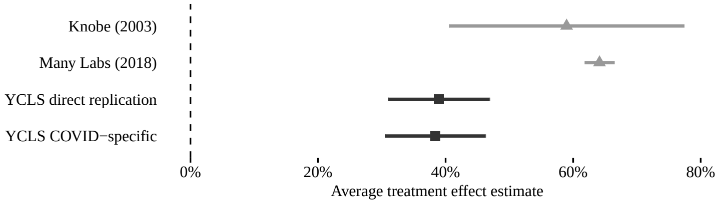
Effect of gain vs. loss frame with party endorsement



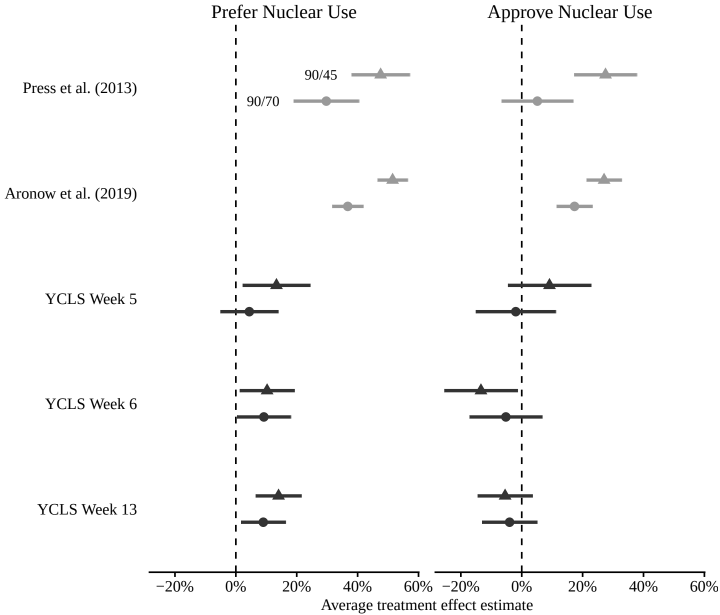
Effect of policy-specific information on support for foreign aid



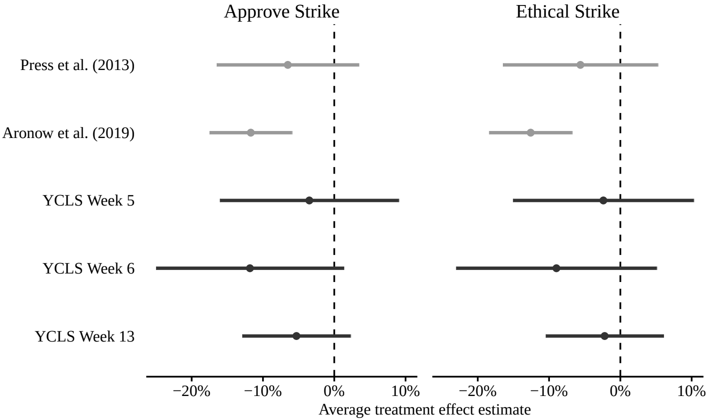
Effect of “Harm” vs. “Help” frame on perceived intentionality



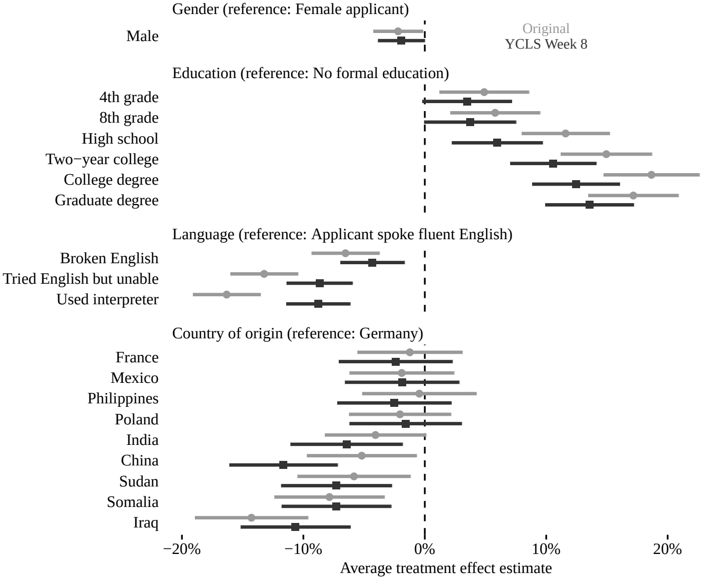
Effects of relative success of "nuclear" v. "conventional" attack on support for U.S. strike on Al-Qaeda weapons lab in Syria

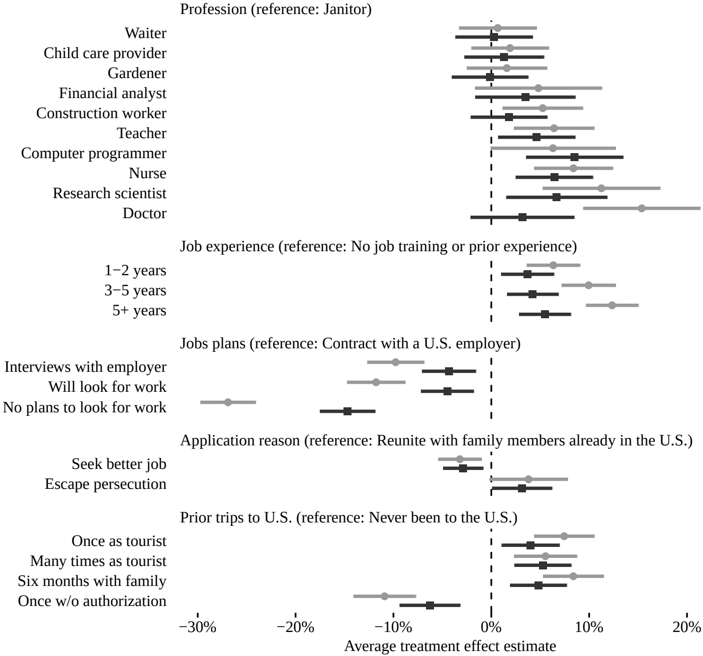


Effects of “nuclear” v. “conventional” attack on support for retrospective U.S. strike on Al-Qaeda nuke lab in Syria

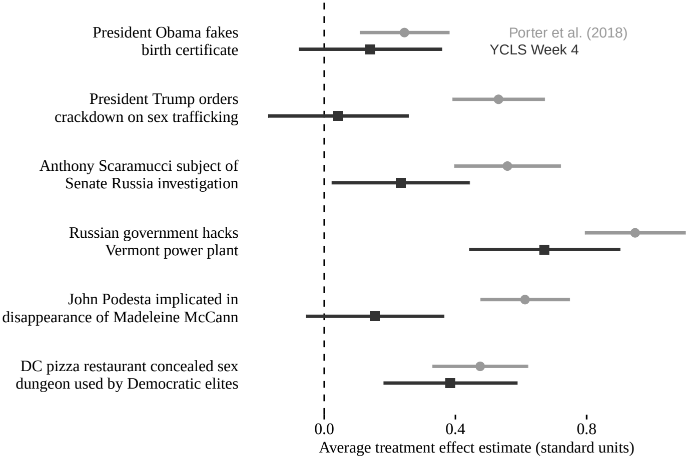


Effects of immigrant attributes on support for admission to U.S.

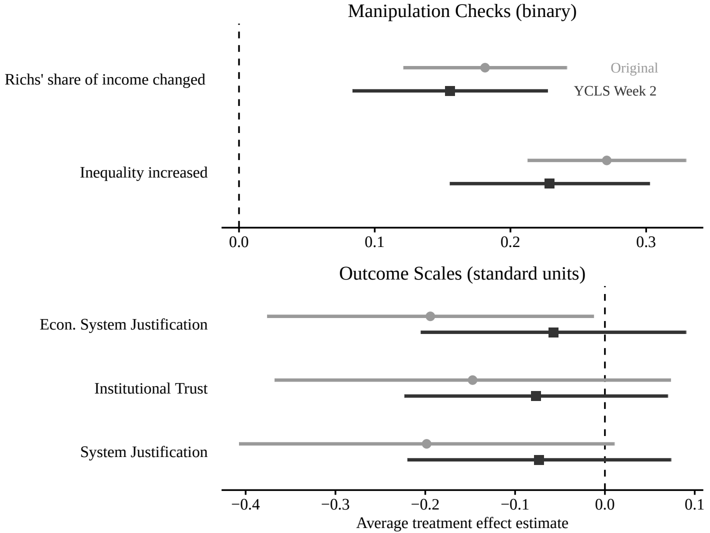



Profession (reference: Janitor)
Waiter
Child care provider
Gardener
Financial analyst
Construction worker
Teacher
Computer programmer
Nurse
Research scientist
Doctor
Job experience (reference: No job training or prior experience)
1−2 years
3−5 years
5+ years
Jobs plans (reference: Contract with a U.S. employer)
Interviews with employer
Will look for work
No plans to look for work
Application reason (reference: Reunite with family members already in the U.S.)
Seek better job
Escape persecution
Prior trips to U.S. (reference: Never been to the U.S.)
Once as tourist
Many times as tourist
Six months with family
Once w/o authorization
−30%
−20%
−10%
0%
10%
20%
Average treatment effect estimate

Effects of "fake news" corrections on disagreement with inaccurate statements



Effects of "high inequality" treatment on comprehension questions and system justification scales



Effect of corruption perceptions on trust in government and support for redistribution