Experiments vs the rest?

William Lowe

Hertie School of Governance

13th September 2020

Experiments vs the rest?

An old but popular view:

Shot: Randomization (and randomized controlled trials) are the gold standard for causal inference. Everything else is

- → at best quasi-experiment
- → at worst description

Chaser: RCTs lack external validity, so we never know whether they generalize

Experiments vs the rest?

An old but popular view:

Shot: Randomization (and randomized controlled trials) are the gold standard for causal inference. Everything else is

- → at best quasi-experiment
- → at worst description

Chaser: RCTs lack external validity, so we never know whether they generalize

What we'll argue here:

Shot: Randomization and RCTs are great, but as soon as they go wrong, or we want to generalize them, we'll need all the tools from observational causal inference

Chaser: RCTs lack external validity. That's why we like them.

Causal effects

Why so serious (about experiments)?

An operational equivalence:

→ The change in Y when you step into a system to change X

→ the change in Y when you randomise X in a (large enough) experiment

Causal effects

Why so serious (about experiments)?

An operational equivalence:

- → The change in Y when you step into a system to change X
- → the change in Y when you randomise X in a (large enough) experiment

Some types:

- → Lab experiments
- → Field experiments
- → 'Natural' experiments

in rough order of

- → how seriously people take them
- → how hard they are to analyze

A field experiment

Gerber et al. (2008) tried to get eligible voters in New Haven to actually vote by sending them postcards

Past attempts:

- → telephone calls
- → personal visits

A field experiment

Gerber et al. (2008) tried to get eligible voters in New Haven to actually vote by sending them postcards

Past attempts:

- telephone calls
- personal visits

Each of four postcard messages was a *randomized* treatment X for about 20,000 households

- Voting is your civic duty
- You are being studied (by us)
- We know whether you voted last time
- Your neighbours will know too, but let us tell you about them

Outcome Y was voting in the primary.

Results

People care what their neighbours think (who knew?)

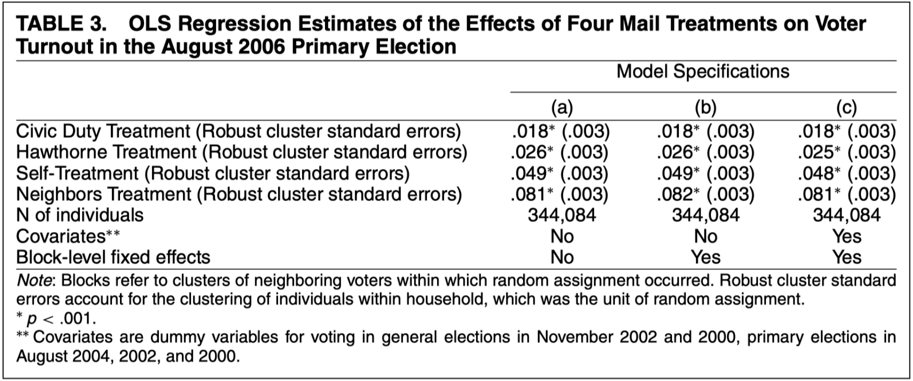


TABLE 3. OLS Regression Estimates of the Effects of Four Mail Treatments on Voter Turnout in the August 2006 Primary Election

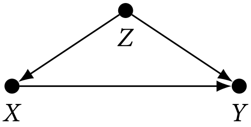
	Model Specifications		
	(a)	(b)	(c)
Civic Duty Treatment (Robust cluster standard errors)	.018* (.003)	.018* (.003)	.018* (.003)
Hawthorne Treatment (Robust cluster standard errors)	.026* (.003)	.026* (.003)	.025* (.003)
Self-Treatment (Robust cluster standard errors)	.049* (.003)	.049* (.003)	.048* (.003)
Neighbors Treatment (Robust cluster standard errors)	.081* (.003)	.082* (.003)	.081* (.003)
N of individuals	344,084	344,084	344,084
Covariates**	No	No	Yes
Block-level fixed effects	No	Yes	Yes

Note: Blocks refer to clusters of neighboring voters within which random assignment occurred. Robust cluster standard errors account for the clustering of individuals within household, which was the unit of random assignment.

* $p < .001$.

** Covariates are dummy variables for voting in general elections in November 2002 and 2000, primary elections in August 2004, 2002, and 2000.

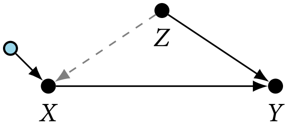
Randomization



Here $Y^{X=x}$ and X both depend on Z, so

$$Y^{X=0}, Y^{X=1} \not\perp\!\!\!\perp X$$

because they share a common factor, e.g. Z is political party membership



But here only $Y^{X=x}$ depends on Z, so

$$Y^{X=0}, Y^{X=1} \perp\!\!\!\perp X$$

Randomization

In principle randomizing is *sufficient* to identify the effect of X on Y

Why bother to also control for stuff?

Randomization

In principle randomizing is *sufficient* to identify the effect of X on Y

Why bother to also control for stuff?

→ Precision

Unlike in observation studies with confounding, this is not necessary for identification

→ But conditioning doesn't know *why* you're doing it, so the process is the same

AND CONTROL

Consider an randomized experiment with m subjects with $X = 1$ and $N - m$ subjects with $X = 0$.

AND CONTROL

Consider an randomized experiment with m subjects with $X = 1$ and $N - m$ subjects with $X = 0$. From *nature's* standpoint

$$\begin{aligned} \text{ATE} &= \text{E}\left[Y^{X=1} \mid X = 1\right] - \text{E}\left[Y^{X=0} \mid X = 0\right] \\ &= \frac{1}{N}\sum_{i}^{N} Y_i^{X=1} - \frac{1}{N}\sum_{i}^{N} Y_i^{X=0} \end{aligned}$$

will have variance

$$\text{Var}(\text{ATE}) = \frac{1}{N-1}\left[\frac{m\ \text{Var}(Y_i^{X=0})}{N-m} + \frac{(N-m)\ \text{Var}(Y_i^{X=1})}{m} + 2\text{Cov}(Y_i^{X=0}, Y_i^{X=1})\right]$$

AND CONTROL

When is this *smaller*?

$$\text{Var}(\text{ATE}) = \frac{1}{N-1}\left[\frac{m\ \text{Var}(Y_i^{X=0})}{N-m} + \frac{(N-m)\ \text{Var}(Y_i^{X=1})}{m} + 2\text{Cov}(Y_i^{X=0}, Y_i^{X=1})\right]$$

Larger N

→ Run a large experiment

Smaller $\text{Var}(Y_i^{X=1})$ and $\text{Var}(Y_i^{X=0})$

→ Block into homogenous groups or use good predictors.

→ Put proportionally more subjects in the noisier condition

Smaller $\text{Cov}(Y_i^{X=0}, Y_i^{X=1})$

→ Sadly you can’t do much about this. Best case: a *negative* covariance

SAY WHAT?

What would negative $\text{Cov}(Y_i^{X=0}, Y_i^{X=1})$ be?

If δ_i is the treatment effect on subject i then $Y_i^{X=1} = Y_i^{X=0} + \delta_i$, then

$$\begin{aligned}\text{Cov}(Y_i^{X=0}, Y_i^{X=1}) &= \text{Cov}(Y_i^{X=0}, Y_i^{X=0} + \delta_i) \\ &= \text{Var}(Y_i^{X=0}) + \text{Cov}(Y_i^{X=0}, \delta_i)\end{aligned}$$

is negative when treatment effects are biggest for those subjects with the lowest expected untreated outcomes

(when $\delta_i = \delta$ they're perfectly positively correlated)

Control: blocking or conditioning

If we believe that potential outcomes are going to vary according to things we can measure, say Z, we can block on that (or those):

- → Divide up Z
- → Randomize X *within* levels of Z

Or we can run the experiment first, then analyze it conditioning on Z, e.g. with regression

Either way:

- → More precision in estimating $Y^{X=0}$ and/or $Y^{X=1}$, then more precision for the ATE
- → Not always (Freedman, 2008), but mostly (Lin, 2013)

The *smaller* the experiment (<100 cases), the more that blocking is preferable

- → Removes chance imbalance between X and Z

Getting the vote out

Gerber et al. (2008) do a bit of both

They *block* using the postal route (it's a bit unclear from the paper)

and statistically *control* for

> *a set of known predictors of voting in primaries: turnout history in previous primary and general elections, gender, number of registered voters in the household, and age.*

Blocking practicalities

How to block, if you have subjects but haven't run the experiment yet?

→ Informal and manual: Think of what you'd put in a regression model block on those variables

→ Automated: Use *matching* technology to choose groups, (e.g. Moore, 2012, and the blockTools package)

This is a slightly ironic use of matching, since matching normally takes non-experimental data and tries to make it like a randomized (but not blocked) experiment (see King & Nielsen, 2019, later in the course).

Just say yes to control

So, you should block, or control for (post-treatment) things, or both.

Estimator precision was a *non-causal inference* reason to get out the regression tools

Let’s see some causal inference reasons to do so…

Just say yes to control

So, you should block, or control for (post-treatment) things, or both.

Estimator precision was a *non-causal inference* reason to get out the regression tools

Let's see some causal inference reasons to do so…

Reminder: it's seldom a good idea to control for things caused by treatment

Things fall apart

In many experimental situations, people don't (or can't) 'comply' with their treatment assignments

- → You are assigned to $X = 1$ (be treated) but you $X = 0$ (didn't), a.k.a. 'failure to treat'
- → You are assigned to $X = 0$ (not be treated) but you $X = 1$ (get treated)

When there is *only* failure to treat, this is

- → one-sided non compliance

When both happen this is

- → two-sided non compliance

Things fall apart

Sometimes you *expect* one-sided non compliance in an 'encouragement design', e.g.

- → invitations, coupons, cheques in the mail

particularly when it would be unethical to coerce

In the vote experiment

- → You could miss the postcard in your stack of junk mail, or
- → not read it because it looked like yet another get out the vote study
- → The postal service could lose or delay it

Things fall apart

Sometimes you *expect* one-sided non compliance in an ‘encouragement design’, e.g.

- invitations, coupons, cheques in the mail

particularly when it would be unethical to coerce

In the vote experiment

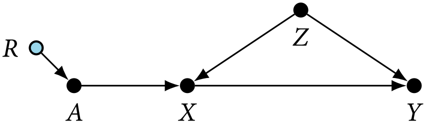
- You could miss the postcard in your stack of junk mail, or
- not read it because it looked like yet another get out the vote study
- The postal service could lose or delay it

For much policy work, you also expect one-sided non compliance

- You can change a law, not everyone will follow it
- Worse, if you do the change may have *other* effects on the outcome you care about

Things fall apart

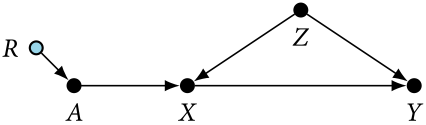
Non-compliance has broken our experiment



because we are randomizing treatment assignment A, not treatment X.

Things fall apart

Non-compliance has broken our experiment



because we are randomizing treatment assignment A, not treatment X.

So what to do with one-sided non-compliance?

If we never know, we can't do much. But let's assume we *know* whether treatment was actually taken.

Things fall apart

Some natural options:

1. Compare those *assigned* to treatment with those *assigned* to control
2. Compare those who *actually* got treated to those assigned to control (and definitely untreated)
3. Compare the actually treated to everyone else

Things fall apart

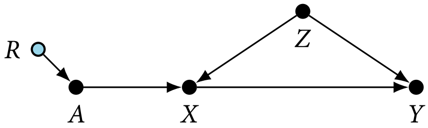
Some natural options:

1. Compare those *assigned* to treatment with those *assigned* to control
2. Compare those who *actually* got treated to those assigned to control (and definitely untreated)
3. Compare the actually treated to everyone else

None of these are good

- → Option 1 successfully answers a different question. (But maybe we like that question!)
- → Option 2 and 3 recreate an *observational study*. If there are common causes of not taking treatment that also affect outcomes, it is confounded

How to think about one-sided non-compliance



Our treatment variable X is now $X_i^{A=1} = 1$ if case i was assigned to treatment and got treated, and $X_i^{A=1} = 0$ when they were assigned to treatment, but *didn't* get treated

We are thinking about one-sided compliance so we know that $X_i^{A=0}$ is always 0.

One-sided non-compliance

We can now define two *types* of subject

$$\begin{aligned}\text{Complier}: &\quad X_i^{A=1} = 1 \;\text{ and }\; X_i^{A=0} = 0\\ \text{Never taker}: &\quad X_i^{A=1} = 0 \;\text{ and }\; X_i^{A=0} = 0\end{aligned}$$

We can't really know who is in which group, however we can see the consequences

Let's revisit our options

The options

Some natural options:

1. Compare those *assigned* to treatment with those *assigned* to control
   - → (Compliers + Never takers) vs (Compliers + Never takers)
2. Compare those who *actually* got treated to those assigned to control (and definitely untreated)
   - → Compliers vs (Compliers + Never takers)
3. Compare the actually treated to everyone else
   - → Compliers vs (Compliers + Never takers)

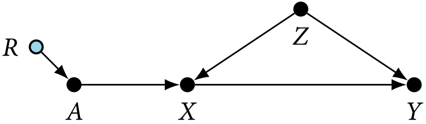
It’s not hard to imagine that Compliers are *not really comparable* to Never takers

Effects

It's useful here to define some new causal effects

Option 1 estimates the *Intention to Treat* effect. Actually there are two, one for X and one for Y:

$$\text{ITT}_X = \text{E}\left[X^{A=1} - X^{A=0}\right]$$
$$\text{ITT} = \text{E}\left[Y^{A=1} - Y^{A=0}\right]$$



Here, the effect of A on X is the ITT_X and the total effect of A on Y is the ITT (Option 1).

But how to get the effect of X on Y?

Effects

Our other options don't compare anything particularly helpful.

The best we can ask for is the *Complier Average Treatment* effect

$$\text{CATE} = \mathrm{E}\left[Y^{X=1} - Y^{X=0} \mid X^{A=1} = 1\right]$$

The overall ATE is a weighted average of this and the ATE for Never takers

→ but we don't know the weights!

Effects

From the graph, you perhaps noticed that

- A is an instrument for X
- CATE is a *Local Average Treatment Effect* (LATE)
- We need an exclusion restriction to estimate it

Exclusion restriction:

- Assignment A does not affect outcomes Y except through treatment X
- Equivalently: no $A \longrightarrow Y$ arrow in the graph

Then

$$\text{CATE} \;=\; \mathrm{E}[\,Y^{X=1} - Y^{X=0} \mid X^{A=1} = 1] \;=\; \frac{\text{ITT}}{\text{ITT}_X}$$

Two sided non-compliance

New problems, new people:

$$\begin{aligned} \text{Always taker}: &\quad X_i^{A=1} = 1 \text{ and } X_i^{A=0} = 1 \\ \text{Complier}: &\quad X_i^{A=1} = 1 \text{ and } X_i^{A=0} = 0 \\ \text{Defier}: &\quad X_i^{A=1} = 0 \text{ and } X_i^{A=0} = 1 \\ \text{Never taker}: &\quad X_i^{A=1} = 0 \text{ and } X_i^{A=0} = 0 \end{aligned}$$

Since we don't know the proportions of each type, it seems like *anything* can happen (and in theory it can)

→ We'll need more assumptions. A standard one is *monotonicity*: There are no Defiers

Then we can estimate CATE as before (The new Always takers don't affect the estimation of ITT or ITT_X)

Compliance

Non-compliance happens, maybe even by design

When it does, we're in the common situation of having only partial control over how the experiment goes

Principle:

→ You can't always randomize the thing you want, but sometimes you can randomize a thing you need

We'll revisit instrumental variable analysis in more detail later in the course

Compliance

Non-compliance happens, maybe even by design

When it does, we're in the common situation of having only partial control over how the experiment goes

Principle:

→ You can't always randomize the thing you want, but sometimes you can randomize a thing you need

We'll revisit instrumental variable analysis in more detail later in the course

Let's turn to a common criticism of even the biggest, most vigorously randomized, beautifully compliant, and superbly controlled studies

→ External validity

External validity

> *'External validity' asks the question of generalizability: To what populations, settings, treatment variables, and measurement variables can this effect be generalized?*
>
> *(Shadish et al., 2002)*

> *An experiment is said to have "external validity" if the distribution of outcomes realized by a treatment group is the same as the distribution of outcome that would be realized in an actual program.*
>
> *(Manski, 2008)*

> *Extrapolation across studies requires some understanding of the reasons for the differences.*
>
> *(Cox, 1958)*

External validity

Recall the ATE (implicitly) averages over subgroup ATEs, e.g.

- → the average of the ATT and the ATC, weighted by the treatment proportion
- → the average of the effect for men and the effect for women, weighted by the gender distribution
- → the average over combinations of "previous primary and general elections, gender, number of registered voters in the household, and age"

more generally, the weighted average of the causal effects for each value of Z weighted by the marginal distribution of $P(Z)$

In *experiments* we sometimes want to learn about subgroups, so focus on one subgroup, say $Z = 1$

In *observational* research we need to average over the confounders explicitly, e.g. in the 'adjustment formula'

External validity

We sometimes hear dark warnings about generalizing effects to new populations

- → This *can* work when the effect is constant
- → But usually not when if the effect differs by group, because group distributions may also differ

How do we transport them?

Effect transportation

Populations can differ by

- → Propensity to be treated (by X)
- → Their distribution of subgroups (by Z)

If we are learning about the causal effect of X then we actually don't need to worry about the distribution of X in the new population

- → The causal effect is conditional on X by definition

Effect transportation

Populations can differ by

→ Propensity to be treated (by X)

→ Their distribution of subgroups (by Z)

If we are learning about the causal effect of X then we actually don't need to worry about the distribution of X in the new population

→ The causal effect is conditional on X by definition

What we *do* need to worry about is the new subgroup distribution Z^* when $P(Z) \neq P(Z^*)$

→ but we can just go measure that

So to infer the effect on the new population we can average the subgroups again, but *weighted by* $P(Z^*)$ instead of $P(Z)$ (Bareinboim & Pearl, 2016)

Effect transportation

> *Differences in propensity to receive treatment do not matter for transportability of causal effects. What matters are potential effect-modifiers.*
>
> *(Cinelli & Bareinboim, 2019)*

See also

→ Rothman et al. (2013) ‘Why representativeness should be avoided’

→ Harrell (2020) ‘Implications of interactions in treatment comparisons’ (interactions because group-specific ATEs are estimated using $X \times Z$ interactions in regression models)

The RCT is just the beginning

> *Unless one wants to confine experimental results to the strict conditions of the studied sub-population, even with a perfect RCT one still needs to go through a transportability exercise (ie, causal modeling)*
>
> *(Cinelli & Bareinboim, 2019)*

Summing up

Randomized experiments are great but as soon as we

- need more precision
- find that things have gone wrong with treatment assignment
- want to transport our findings to a new population

we must resort to observational causal inference tools

- Instrumental variable analysis
- Regression

If we do it badly, we get purely descriptive, causally uninterpretable comparisons.

References

Bareinboim, E. & Pearl, J. (2016). 'Causal inference and the data-fusion problem'. *Proceedings of the National Academy of Sciences*, *113*(27), 7345–7352.

Cinelli, C. & Bareinboim, E. (2019, September). *Generalizability in causal inference*. University of Caifornia at Riverside.

Cox, D. R. (1958). 'Some problems connected with statistical inference'. *The Annals of Mathematical Statistics*, *29*(2), 357–372.

Freedman, D. A. (2008). 'Randomization does not justify logistic regression'. *Statistical Science*, *23*(2), 237–249.

Gerber, A. S., Green, D. P. & Larimer, C. W. (2008). 'Social pressure and voter turnout: Evidence from a large-scale field experiment'. *American Political Science Review*, *102*(01), 33–48.

King, G. & Nielsen, R. (2019). 'Why propensity scores should not be used for matching'. *Political Analysis*, *27*(4), 435–454.

References

Lin, W. (2013). ‘Agnostic notes on regression adjustments to experimental data: Reexamining freedman’s critique’. *The Annals of Applied Statistics*.

Manski, C. F. (2008). ‘Identification for prediction and decision’. Harvard University Press.

Moore, R. T. (2012). ‘Multivariate continuous blocking to improve political science experiments’. *Political Analysis*, *20*(4), 460–479.

Rothman, K. J., Gallacher, J. E. & Hatch, E. E. (2013). ‘Why representativeness should be avoided’. *International Journal of Epidemiology*, *42*(4), 1012–1014.

Shadish, W. R., Cook, T. D. & Campbell, D. T. (2002). ‘Experimental and quasi-experimental designs for generalized causal inference’. Houghton Mifflin.