1. Potential Outcomes and Randomization

PhD Applied Methods

Duncan Webb
NovaSBE

Spring 2026

Intros

- **Me**: Duncan Webb
  - Development economist (field experiments in India, Madagascar, Colombia)
  - PhD from Paris School of Economics (2024)
  - Email: dmbwebb@gmail.com — Office: B115B
  - Office hours: Wednesdays 1–2pm

Intros

- **Me**: Duncan Webb
  - Development economist (field experiments in India, Madagascar, Colombia)
  - PhD from Paris School of Economics (2024)
  - Email: dmbwebb@gmail.com — Office: B115B
  - Office hours: Wednesdays 1–2pm
- **You**: Name, research interests, what you hope to get from this course

Course Structure

- **Goal**: Deep understanding of modern causal inference methods
- **Format**: One 3-hour class per week
- **Topics**:
  1. Potential Outcomes and Randomization (today)
  2. Instrumental Variables
  3. Difference-in-Differences
  4. Regression Discontinuity Design
  5. Empirical Tools

Course Structure

- **Goal**: Deep understanding of modern causal inference methods
- **Format**: One 3-hour class per week
- **Topics**:
  1. Potential Outcomes and Randomization (today)
  2. Instrumental Variables
  3. Difference-in-Differences
  4. Regression Discontinuity Design
  5. Empirical Tools
- **Approach**: Theory + applications + implementation

Assessment

- **Final exam** (80%): 1.5-hour closed-book exam
    - Theoretical derivations and proofs
    - Applied reasoning and interpretation

Assessment

- **Final exam** (80%): 1.5-hour closed-book exam
  - Theoretical derivations and proofs
  - Applied reasoning and interpretation

- **Problem sets** (20%): 4 problem sets throughout the course
  - Applied/empirical questions (Stata, R, or Python)
  - Mathematical/theoretical questions
  - Individual work
  - Due Tuesdays at 12pm, starting next week
  - Submit to `dmbwebb@gmail.com`: **.tex**, **.pdf**, and **code files**
  - Include your name in the output

Assessment

- **Final exam** (80%): 1.5-hour closed-book exam
  - Theoretical derivations and proofs
  - Applied reasoning and interpretation

- **Problem sets** (20%): 4 problem sets throughout the course
  - Applied/empirical questions (Stata, R, or Python)
  - Mathematical/theoretical questions
  - Individual work
  - Due Tuesdays at 12pm, starting next week
  - Submit to `dmbwebb@gmail.com`: **.tex**, **.pdf**, and **code files**
  - Include your name in the output

- **AI policy**: You may use AI assistants, but 80% of your grade is a closed-book exam — AI only helps if you actually learn

Causality and understanding the world

- “We do not have knowledge of a thing until we have grasped its why, that is to say, its cause.” $\sim$ Aristotle

Causality and understanding the world

- “We do not have knowledge of a thing until we have grasped its why, that is to say, its cause.” $\sim$ Aristotle
- Not all research estimates a causal relationship, but the implication or takeaway of a paper is **almost always** a causal one
- Particularly important for **policy evaluation**:
  - What is the effect of microfinance on consumption?
  - What is the effect of reducing class size on education outcomes?
  - What is the effect of reducing the price of a good on its consumption?

Causality and correlation

- For a very long time, economists made causal claims based on **correlations** and very shaky assumptions
- Until the **credibility revolution** (Angrist and Pischke, 2010) which formalized the conditions under which we could claim causality
  - And the use of **randomization** (or quasi-random events) to make those claims
- Goal for this class: Deep understanding of the tools we can use to make causal claims

Rubin’s potential outcomes framework

- The **potential outcomes framework** gives us a precise framework for thinking about when we can correctly claim to estimate the causal effect of some treatment

- The goal is to estimate the **causal effect** of some treatment, e.g.,
  - “Small class at school”
  - “Job training program”
  - ...

Rubin’s potential outcomes framework

- The **potential outcomes framework** gives us a precise framework for thinking about when we can correctly claim to estimate the causal effect of some treatment

- The goal is to estimate the **causal effect** of some treatment, e.g.,
  - “Small class at school”
  - “Job training program”
  - ...

- Note that you can also have multiple treatments (e.g., small class, medium class, large class) and continuous treatments (University fees), but we’ll get to that later

Counterfactuals

- **Outcome measure** is the outcome we care about
- Let Y_i be the observed outcome for individual i
- For example:
  - i's wages in adulthood when examining impact of a job training program
  - i's test scores at school (for class size)
  - How much i discriminates against a minority (for prejudice-reduction intervention)

Counterfactuals

For each person *i* we assume there are two potential outcomes:

- $Y_i(1)$ is the outcome we would observe **if** she received the treatment
- $Y_i(0)$ is the outcome we would observe **if** she did not receive the treatment

We can compactly write this by defining $D_i \in \{0, 1\}$ as the treatment status of individual *i*, and the counterfactuals are $Y_i(D_i)$

Counterfactuals

For each person i we assume there are two potential outcomes:

- $Y_i(1)$ is the outcome we would observe **if** she received the treatment
- $Y_i(0)$ is the outcome we would observe **if** she did not receive the treatment

We can compactly write this by defining $D_i \in \{0, 1\}$ as the treatment status of individual i, and the counterfactuals are $Y_i(D_i)$

If some individuals get the treatment, and some don't... **what can we observe?**

Counterfactuals

- If i is not treated $(D_i = 0)$ then we **only** observe $Y_i(0)$ and $Y_i(1)$ is an unobserved counterfactual
- If i is not treated $(D_i = 1)$ then we **only** observe $Y_i(1)$ and $Y_i(0)$ is an unobserved counterfactual

Counterfactual quiz

Let’s say our “treatment” (D_i) is a **job training program**:

- What is the observed counterfactual for someone in the job training?

Counterfactual quiz

Let’s say our “treatment” (D_i) is a **job training program**:

- What is the observed counterfactual for someone in the job training?
- What is $Y_i(0)$ for someone in the control group?

Counterfactual quiz

Let’s say our “treatment” (D_i) is a **job training program**:

- What is the observed counterfactual for someone in the job training?
- What is $Y_i(0)$ for someone in the control group?
- What is $Y_i(0)$ for someone in the job training?

Counterfactual quiz

Let’s say our “treatment” (D_i) is a **job training program**:

- What is the observed counterfactual for someone in the job training?
- What is $Y_i(0)$ for someone in the control group?
- What is $Y_i(0)$ for someone in the job training?
- Can we observe $Y_i(1)$ for someone who doesn’t get the training?

Counterfactual quiz

Let's say our "treatment" (D_i) is a **job training program**:

- What is the observed counterfactual for someone in the job training?
- What is $Y_i(0)$ for someone in the control group?
- What is $Y_i(0)$ for someone in the job training?
- Can we observe $Y_i(1)$ for someone who doesn't get the training?
- What are $Y_i(0)$ and $Y_i(1)$ for someone who isn't even in the data?

Causal effects

Using this framework, how would we write the **causal effect of the treatment on individual** i?

Causal effects

Using this framework, how would we write the **causal effect of the treatment on individual** i?

$$\Delta_i := Y_i(1) - Y_i(0) \tag{1}$$

This is the main thing we are trying to estimate!

In general, $Y_i(1)$ and $Y_i(0)$ can be different across people, and so Δ_i may be different for each person too ("heterogeneous treatment effects")

Fundamental identification problem

Question: What is the fundamental difficulty with estimating Δ_i, the causal effect of the treatment on individual i?

Fundamental identification problem

Question: What is the fundamental difficulty with estimating Δ_i, the causal effect of the treatment on individual i?

Answer: For a specific person i, we do not and **cannot even principle** observe both $Y_i(1)$ and $Y_i(0)$.

We do not know what exactly what would have happened to Donald Trump (and the world) if he had not been shot, because in fact he was. So we cannot know for sure the causal effect of him being shot.

Fundamental identification problem

Question: What is the fundamental difficulty with estimating Δ_i, the causal effect of the treatment on individual i?

Answer: For a specific person i, we do not and **cannot even principle** observe both $Y_i(1)$ and $Y_i(0)$.

We do not know what exactly what would have happened to Donald Trump (and the world) if he had not been shot, because in fact he was. So we cannot know for sure the causal effect of him being shot.

This is called the **fundamental identification problem**.

Some assumptions

What are some assumptions built into my stipulation that there are some values $Y_i(1), Y_i(0)$?

1. **Partial equilibrium**: these counterfactuals are implicitly defined in a given environment, e.g., what would have happened to Homer *if* he got a job training program, but holding fixed the macroeconomic situation

Some assumptions

What are some assumptions built into my stipulation that there are some values $Y_i(1), Y_i(0)$?

1. **Partial equilibrium**: these counterfactuals are implicitly defined in a given environment, e.g., what would have happened to Homer *if* he got a job training program, but holding fixed the macroeconomic situation
2. **Treatment doesn't affect treated** ("Stable Unit Treatment Value Assumption"): if Homer gets a job training program, that doesn't affect the counterfactuals of Marge or Ned Flanders
   - So this rules out general equilibrium effects, externalities, Hawthorne effects, etc.

Heterogeneous treatment effects

Call $D_i = 0$ if untreated and $D_i = 1$ if treated

Because the effect can be heterogenous, many evaluation parameters. In particular:

$$ATE := \mathbb{E}[Y_i(1) - Y_i(0)] \tag{2}$$

$$ATT := \mathbb{E}[Y_i(1) - Y_i(0)|D_i = 1] \tag{3}$$

- **ATE**: Average treatment effect - all population
- **ATT**: Average treatment on the treated - treated only (for instance, weak students are treated first)

Observed outcomes

How do we compactly write the **actually observed outcome**?

Observed outcomes

How do we compactly write the **actually observed outcome**?

$$Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0) \tag{4}$$

Observed outcomes

How do we compactly write the **actually observed outcome**?

$$Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0) \tag{4}$$

If $D_i = 1$ then we observe $Y_i(1)$
If $D_i = 0$ then we observe $Y_i(0)$
Think of it as a “binary switch”

NB: We can equivalently write this as $Y_i = Y_i(0) + D_i(Y_i(1) - Y_i(0)) = Y_i(0) + D_i\Delta_i$ (i.e., in terms of the effect)

Second quiz

- What is the observed outcome of $D_i = 1$?

Second quiz

- What is the observed outcome of $D_i = 1$?
- What is the unobserved counterfactual of $D_i = 0$?

Regression?

Consider the regression

$$Y_i = \alpha + D_i\beta + u_i \tag{5}$$

Express the counterfactuals $Y_i(0)$ and $Y_i(1)$ that generate this model

Regression?

What is OLS estimator of β in this regression?

$$Y_i = \alpha + D_i\beta + u_i \tag{6}$$

Regression?

OLS estimator is the difference in empirical means

It makes sense as, under $\mathbb{E}[u_i|D_i] = 0$:

$$\mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0] = \alpha + \beta - \alpha + \mathbb{E}[u_i|D_i = 1] - \mathbb{E}[u_i|D_i = 0] = \beta \qquad (7)$$

More general regression model:

$$Y_i(0) = g_0(X_i) + u_0 \tag{8}$$

$$Y_i(1) = g_1(X_i) + u_1 \tag{9}$$

Recall

$$Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0) \tag{10}$$

Then:

$$Y_i = g_0(X_i) + D_i[g_1(X_i) - g_0(X_i) + u_1 - u_0] + u_0 \tag{11}$$

Heterogeneous treatment effect

$$Y_i = g_0(X_i) + D_i \underbrace{[g_1(X_i) - g_0(X_i) + u_1 - u_0]} + u_0 \tag{12}$$

$[g_1(X_i) - g_0(X_i) + u_1 - u_0]$: parameter specific to each individual

In particular:

$$ATE = \mathbb{E}[g_1(X_i) - g_0(X_i) + u_1 - u_0] \tag{13}$$

$$TT = \mathbb{E}[g_1(X_i) - g_0(X_i) + u_1 - u_0 | D_i = 1] \tag{14}$$

If treated and untreated do not have the same distributions for x or $(u_1 - u_0)$, then in general: $TT \neq ATE$

Does OLS of Y on D_i estimate any of those parameters?

$$ATE = \mathbb{E}[g_1(X_i) - g_0(X_i) + u_1 - u_0] \tag{15}$$

$$TT = \mathbb{E}[g_1(X_i) - g_0(X_i) + u_1 - u_0 | D_i = 1] \tag{16}$$

OLS estimates the following:

$$\mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0] = \mathbb{E}[g_1(X_i) + u_1|D_i = 1] - \mathbb{E}[g_0(X_i) + u_0|D_i = 0] \tag{17}$$

Even if $\mathbb{E}[u_k|D_i] = 0, k = 0, 1$, OLS estimates neither one (but it is ATE, and TT, if X_i is distributed equally in the two groups)

Restrictions

$u_1 = u_0$ and $g_1(X_i) = g_0(X_i) + \beta$

lead to a homogenous effects model:

$$Y_i = g_0(X_i) + D_i\beta + u_0 \tag{18}$$

and $TT = ATE = \beta$

NB: in that case, $Y_i(0)$ and $Y_i(1)$ may be heterogenous but $(Y_i(1) - Y_i(0)) = \beta$ homogenous

Outline

1. Potential outcomes framework

2. Selection bias

3. Controlled experiments

4. Clustered randomization

5. Statistical power

6. Conclusion

What do we observe?

Let’s think again about what can we actually observe in the data?

$$\mathbb{E}[Y_i(0)|D_i = 0]? \tag{19}$$

$$\mathbb{E}[Y_i(1)|D_i = 1]? \tag{20}$$

$$\mathbb{E}[Y_i(1)|D_i = 0]? \tag{21}$$

$$\mathbb{E}[Y_i(0)|D_i = 1]? \tag{22}$$

What do we observe?

Identification problem: we “**observe**”

$$\mathbb{E}[Y_i(0)|D_i = 0] \tag{23}$$
$$\mathbb{E}[Y_i(1)|D_i = 1] \tag{24}$$

but not the counterfactuals

$$\mathbb{E}[Y_i(1)|D_i = 0] \tag{25}$$
$$\mathbb{E}[Y_i(0)|D_i = 1] \tag{26}$$

But what would we need to estimate the ATT, for instance?

What do we observe?

Identification problem: we "**observe**"

$$\mathbb{E}[Y_i(0)|D_i = 0] \tag{23}$$

$$\mathbb{E}[Y_i(1)|D_i = 1] \tag{24}$$

but not the counterfactuals

$$\mathbb{E}[Y_i(1)|D_i = 0] \tag{25}$$

$$\mathbb{E}[Y_i(0)|D_i = 1] \tag{26}$$

But what would we need to estimate the *ATT*, for instance?

$$TT = \overbrace{\mathbb{E}[Y_i(1)|D_i = 1]}^{\text{“observed”}} - \overbrace{\mathbb{E}[Y_i(0)|D_i = 1]}^{\text{“unobserved”}} \tag{27}$$

Hypothesis to identify TT:

$$\boxed{\mathbb{E}[Y_i(0)|D_i = 1] = \mathbb{E}[Y_i(0)|D_i = 0] = \mathbb{E}[Y_i(0)]}$$

i.e. no selectivity: treated “are like” untreated

Then

$$ATT = \overbrace{\mathbb{E}[Y_i(1)|D_i = 1]}^{\text{“observed”}} - \overbrace{\mathbb{E}[Y_i(0)|D_i = 1]}^{\text{“unobserved”}} \tag{28}$$

$$= \overbrace{\mathbb{E}[Y_i(1)|D_i = 1]}^{\text{“observed”}} - \overbrace{\mathbb{E}[Y_i(0)|D_i = 0]}^{\text{“observed”}} \tag{29}$$

In this very simple case, compare empirical means in each group

The counterfactual for a group is simply the observed outcome of the other group

NB: to identify *ATE* we need more:

$$\boxed{\mathbb{E}[Y_i(0)|D_i] = \mathbb{E}[Y_i(0)]} \text{ and } \boxed{\mathbb{E}[Y_i(1)|D_i] = \mathbb{E}[Y_i(1)]}$$

$$ATE = \overbrace{\mathbb{E}[Y_i(1)]}^{\text{“unobserved”}} - \overbrace{\mathbb{E}[Y_i(0)]}^{\text{“unobserved”}} \tag{30}$$

$$= \overbrace{\mathbb{E}[Y_i(1)|D_i = 1]}^{\text{“observed”}} - \overbrace{\mathbb{E}[Y_i(0)|D_i = 0]}^{\text{“observed”}} \tag{31}$$

Under these assumptions, we also identify the *ATT* - **why**?

NB: to identify *ATE* we need more:

$$\mathbb{E}[Y_i(0)|D_i] = \mathbb{E}[Y_i(0)] \text{ and } \mathbb{E}[Y_i(1)|D_i] = \mathbb{E}[Y_i(1)]$$

$$ATE = \overbrace{\mathbb{E}[Y_i(1)]}^{\text{“unobserved”}} - \overbrace{\mathbb{E}[Y_i(0)]}^{\text{“unobserved”}} \tag{30}$$

$$= \overbrace{\mathbb{E}[Y_i(1)|D_i = 1]}^{\text{“observed”}} - \overbrace{\mathbb{E}[Y_i(0)|D_i = 0]}^{\text{“observed”}} \tag{31}$$

Under these assumptions, we also identify the *ATT* **- why**?

$\implies$ Because we assumed that the counterfactuals are similarly distributed (on average) in both populations

But, in general

$$\boxed{\mathbb{E}[Y_i(0)|D_i = 1] \neq \mathbb{E}[Y_i(0)|D_i = 0]}$$

The “naive” estimator (difference in observed means) is then a biased estimator for *ATT*:

$$\begin{aligned}
&\mathbb{E}[Y_i(1)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0] && (32)\\
&= \underbrace{[\mathbb{E}[Y_i(1)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 1]]} + \underbrace{[\mathbb{E}[Y_i(0)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0]]} && (33)\\
&= ATT + \text{Selection Bias} && (34)
\end{aligned}$$

where bias $[\mathbb{E}[Y_i(0)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0]]$ is the difference between (average) counterfactual $Y_i(0)$ in the two populations (treated and untreated)

Why is selection bias quite likely?

Simple **Roy model**: “I am in if this is worth it”

$$D_i = 1 \text{ if } Y_i(1) - Y_i(0) > c$$

Why is selection bias quite likely?

Simple **Roy model**: “I am in if this is worth it”

$$D_i = 1 \text{ if } Y_i(1) - Y_i(0) > c$$

Then, in general

$$\begin{aligned} \mathbb{E}[Y_i(0)|D_i = 1] &= \mathbb{E}[Y_i(0)|Y_i(0) < Y_i(1) - c] && (35) \\ &\neq \mathbb{E}[Y_i(0)|Y_i(0) \geq Y_i(1) - c] = \mathbb{E}[Y_i(0)|D_i = 0] && (36) \end{aligned}$$

In this case, selectivity stems from

- **Comparative advantages** ($Y_i(1) - Y_i(0)$ large for some, small for others). Most simple instance: participants have smaller $Y_i(0)$, thus larger potential gain
- **Heterogeneity in cost** c (if it is correlated with $Y_i(0)$)

Why is selection bias quite likely?

Another reason for selection bias: **administrative rules**

For instance:

- **"Cream-skimming"**: they choose "the best", and $\mathbb{E}[Y_i(0)|D_i = 1] > \mathbb{E}[Y_i(0)|D_i = 0]$
- **Remedial targeting**: e.g. focus on intervening with weak kids in the class, so $\mathbb{E}[Y_i(0)|D_i = 1] < \mathbb{E}[Y_i(0)|D_i = 0]$

Link with endogeneity

Selectivity is similar to **endogeneity** of D_i in a regression

For simplicity, focus on a simple model with homogenous effects, i.e. $\Delta_i = \beta$ for everyone, and $u_1 = u_0 = u_i$:

$$Y_i = \alpha + D_i\beta + u$$

Selectivity is then:

$$\Rightarrow \mathbb{E}[Y_i(0)|D_i = 1] \neq \mathbb{E}[Y_i(0)|D_i = 0] \quad (37)$$

$$\Rightarrow \mathbb{E}[u_i|D_i = 1] \neq \mathbb{E}[u_i|D_i = 0] \quad (38)$$

A least-squares regression would confound treatment effect (β) with the fact that the distribution of u_o is different in treated and untreated populations:

$$\mathbb{E}[Y_i|D_i = 0] = \alpha + \mathbb{E}[u_i|D_i = 0] \tag{39}$$

$$\mathbb{E}[Y_i|D_i = 1] = \alpha + \beta + \mathbb{E}[u_i|D_i = 1] \tag{40}$$

and OLS is an estimator of

$$\mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0] = \beta + (\mathbb{E}[u_i|D_i = 1] - \mathbb{E}[u_i|D_i = 0]) \tag{41}$$

$$= \beta + \text{Bias} \tag{42}$$

Summing up

1. **Counterfactuals** allow us to define the causal effect of a treatment
2. **Identification problem**: some counterfactuals are (always) unobserved
3. **Section bias**: in general, treated outcomes cannot be used to estimate untreated counterfactuals and vice-versa

We have this fundamental identification problem: we can't observe both $Y_i(1)$ and $Y_i(0)$ for an individual i.

We have this fundamental identification problem: we can't observe both $Y_i(1)$ and $Y_i(0)$ for an individual i.

But if we can assume:

$$\mathbb{E}[Y_i(0)|D_i = 1] = \mathbb{E}[Y_i(0)|D_i = 0] = \mathbb{E}[Y_i(0)] \tag{43}$$

we can at least identify the **average** treatment effects.

We have this fundamental identification problem: we can't observe both $Y_i(1)$ and $Y_i(0)$ for an individual i.

But if we can assume:

$$\mathbb{E}[Y_i(0)|D_i = 1] = \mathbb{E}[Y_i(0)|D_i = 0] = \mathbb{E}[Y_i(0)] \tag{43}$$

we can at least identify the **average** treatment effects.

Question: how can we **make this assumption true**?

Outline

1. Potential outcomes framework

2. Selection bias

3. Controlled experiments

4. Clustered randomization

5. Statistical power

6. Conclusion

Idea behind experiments

Simplest way to identify treatment causal effect: make likely the hypotheses

$$\mathbb{E}[Y_i(0)|D_i = 1] = \mathbb{E}[Y_i(0)|D_i = 0] = \mathbb{E}[Y_i(0)] \tag{44}$$
$$\mathbb{E}[Y_i(1)|D_i = 1] = \mathbb{E}[Y_i(1)|D_i = 0] = \mathbb{E}[Y_i(1)] \tag{45}$$

If we draw treated and untreated **randomly** from a population and

$$ATT = \overbrace{\mathbb{E}[Y_i(1)|D_i = 1]}^{\text{“observed”}} - \overbrace{\mathbb{E}[Y_i(0)|D_i = 1]}^{\text{“unobserved”}} \tag{46}$$
$$= \overbrace{\mathbb{E}[Y_i(1)|D_i = 1]}^{\text{“observed”}} - \overbrace{\mathbb{E}[Y_i(0)|D_i = 0]}^{\text{“observed”}} \tag{47}$$

Can be estimated with empirical means (or via regressions)

Idea behind experiments

- **Intuition**: if we randomly select who receives the treatment and who doesn't, then on average it will be similar types of people in each group, and so the average counterfactuals will be the same

- Therefore, any difference we **do** observe after the treatment must be **caused by the treatment**

- This is why randomized controlled trials are called the **gold standard** of evidence (somewhat controversially)

- **Other methods for causal inference** are built on this paradigm – other identification methods "mimic" random assignment into treatment

Use of randomized controlled trials

- **Popularity**: AER+JPE+QJE, 0.8% of published articles in 1983 $\rightarrow$ 8.2% in 2011 (while theory: 58% $\rightarrow$ 19%)
- **Nobel Prize** in economics to Esther Duflo, Abhijit Banerjee, Michael Kremer for pioneering this methodology in development economics.
- **Infrastructure -** organisations like J-PAL and IPA provide infrastructure for this kind of research

Critiques of randomized controlled trials

- **Equilibrium effects -** standard methodology ignores spillover effects or equilibrium effects, although frontier methods and large trials can understand these (see e.g., Egger et al, *Econometrica* 2022)

- **External validity -** how to generalize from the context of the RCT to other contexts

- **Ethics -** is it ethical to deny treatment to the control group? This depends on the context, and what the alternative is, e.g., it's no longer ethical to deny proven medical treatment

- **Mechanisms -** early RCTs tried to measure treatment effects, but high-quality studies now focus a lot on understanding mechanisms using additional treatments or by explicitly testing theory-driven models

Simplest design

Randomize individuals and compare treated and untreated:

$$\bar{Y}_1 - \bar{Y}_0 = \frac{1}{N_1} \sum_{i \in D_i = 1} Y_i \quad - \quad \frac{1}{N_0} \sum_{i \in D_i = 0} Y_i$$

Similar to OLS on

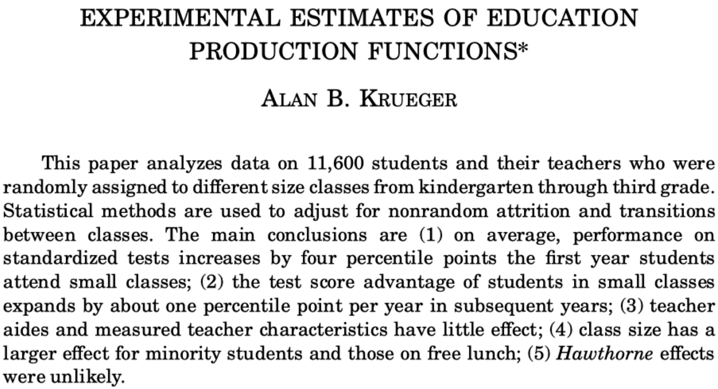
$$y = \alpha + \beta D_i + u$$

(Leave as an exercise to prove it algebraically using OLS matrix formula)

Real life design issues

1. **Balance** between treatment groups in a finite sample
2. **Adding controls**
3. **Imperfect compliance**

Real example: reducing class size in “STAR” program



EXPERIMENTAL ESTIMATES OF EDUCATION PRODUCTION FUNCTIONS*

ALAN B. KRUEGER

This paper analyzes data on 11,600 students and their teachers who were randomly assigned to different size classes from kindergarten through third grade. Statistical methods are used to adjust for nonrandom attrition and transitions between classes. The main conclusions are (1) on average, performance on standardized tests increases by four percentile points the first year students attend small classes; (2) the test score advantage of students in small classes expands by about one percentile point per year in subsequent years; (3) teacher aides and measured teacher characteristics have little effect; (4) class size has a larger effect for minority students and those on free lunch; (5) *Hawthorne* effects were unlikely.

1. **Balance** between treatment groups

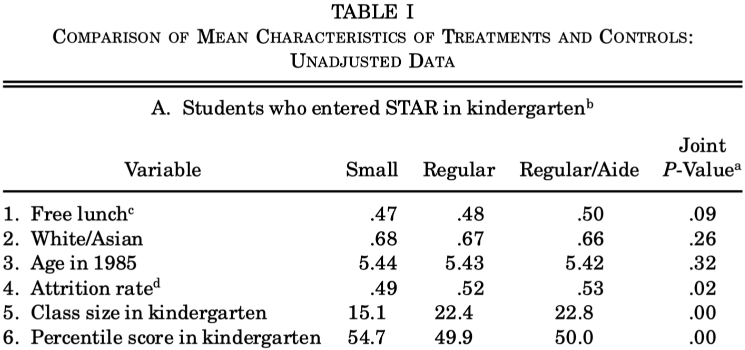


TABLE I
COMPARISON OF MEAN CHARACTERISTICS OF TREATMENTS AND CONTROLS: UNADJUSTED DATA

A. Students who entered STAR in kindergarten[b]

Variable	Small	Regular	Regular/Aide	Joint *P*-Value[a]
1. Free lunch[c]	.47	.48	.50	.09
2. White/Asian	.68	.67	.66	.26
3. Age in 1985	5.44	5.43	5.42	.32
4. Attrition rate[d]	.49	.52	.53	.02
5. Class size in kindergarten	15.1	22.4	22.8	.00
6. Percentile score in kindergarten	54.7	49.9	50.0	.00

2. **Adding controls** with an OLS regression of an RCT

This shows the effect of D_i (“small class”) on Y_i (percentile on standardized test score).

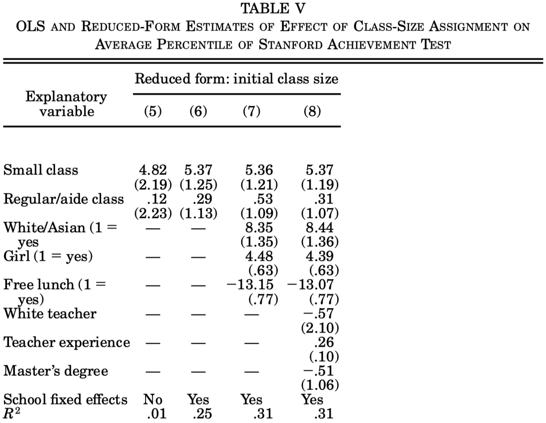


TABLE V
OLS and Reduced-Form Estimates of Effect of Class-Size Assignment on Average Percentile of Stanford Achievement Test

Explanatory variable	Reduced form: initial class size			
	(5)	(6)	(7)	(8)
Small class	4.82	5.37	5.36	5.37
	(2.19)	(1.25)	(1.21)	(1.19)
Regular/aide class	.12	.29	.53	.31
	(2.23)	(1.13)	(1.09)	(1.07)
White/Asian (1 = yes	—	—	8.35	8.44
			(1.35)	(1.36)
Girl (1 = yes)	—	—	4.48	4.39
			(.63)	(.63)
Free lunch (1 = yes)	—	—	−13.15	−13.07
			(.77)	(.77)
White teacher	—	—	—	−.57
				(2.10)
Teacher experience	—	—	—	.26
				(.10)
Master’s degree	—	—	—	−.51
				(1.06)
School fixed effects	No	Yes	Yes	Yes
R^2	.01	.25	.31	.31

Adding controls

OLS often used because it allows you to add controls: **Why?**

Adding controls

OLS often used because it allows you to add controls: **Why?**

If assignment is truly random, conditioning on observed characteristics X_i should not affect point estimates

In effect, we have both $\mathbb{E}[Y_i(0)|D_i] = \mathbb{E}[Y_i(0)]$, and $\mathbb{E}[X_i|D_i] = \mathbb{E}[X_i]$

Therefore X_i and D_i are (mean-)independant and OLS estimation

$$Y_i = X_i\gamma + \beta D_i + u$$

gives rise (asymptotically) to same coefficient as separate OLS

$$Y_i = \beta D_i + u_i'$$

and

$$Y_i = X_i\gamma + u_i''$$

Adding controls

But it’s still useful because it increases precision: **Why?**

Adding controls

But it's still useful because it increases precision: **Why?**

$$V(\hat{\beta}) = \sigma^2(X'X)^{-1}$$

where σ^2 is residual variance

$$y = \beta D_i + u' \tag{48}$$

$$y = \beta D_i + x\gamma + u \tag{49}$$

$V(u') = V(x\gamma) + V(u) > V(u)$

Thus, the second equation estimates the same β but with more precision

Depends on how much X_i explain Y_i (and may **not hold in finite samples**)

Controls and precision

This shows the effect of D_i (“small class”) on Y_i (percentile on standardized test score).

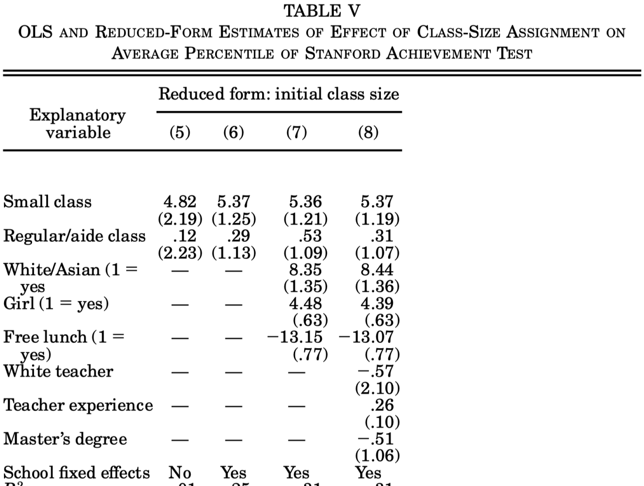


TABLE V

OLS and Reduced-Form Estimates of Effect of Class-Size Assignment on Average Percentile of Stanford Achievement Test

Explanatory variable	Reduced form: initial class size			
	(5)	(6)	(7)	(8)
Small class	4.82 (2.19)	5.37 (1.25)	5.36 (1.21)	5.37 (1.19)
Regular/aide class	.12 (2.23)	.29 (1.13)	.53 (1.09)	.31 (1.07)
White/Asian (1 = yes	—	—	8.35 (1.35)	8.44 (1.36)
Girl (1 = yes)	—	—	4.48 (.63)	4.39 (.63)
Free lunch (1 = yes)	—	—	−13.15 (.77)	−13.07 (.77)
White teacher	—	—	—	−.57 (2.10)
Teacher experience	—	—	—	.26 (.10)
Master’s degree	—	—	—	−.51 (1.06)
School fixed effects	No	Yes	Yes	Yes

Stratified randomization (block randomization)

Motivation: With simple randomization, we might randomly get imbalances on important characteristics (especially in small samples). Can we do better?

Stratified randomization (block randomization)

Motivation: With simple randomization, we might randomly get imbalances on important characteristics (especially in small samples). Can we do better?

Simple randomization: Randomly assign all units to treatment or control

Stratified randomization (block randomization)

Motivation: With simple randomization, we might randomly get imbalances on important characteristics (especially in small samples). Can we do better?

Simple randomization: Randomly assign all units to treatment or control

Stratified randomization: Divide sample into strata (blocks) based on pre-treatment characteristics, then randomize **within each stratum**

Stratified randomization (block randomization)

Motivation: With simple randomization, we might randomly get imbalances on important characteristics (especially in small samples). Can we do better?

Simple randomization: Randomly assign all units to treatment or control

Stratified randomization: Divide sample into strata (blocks) based on pre-treatment characteristics, then randomize **within each stratum**

Example: Study of job training program

- Stratify by: education level (high school vs. college) and gender
- Creates 4 strata: HS males, HS females, College males, College females
- Within each stratum, randomly assign 50% to treatment, 50% to control

Stratified randomization (block randomization)

Motivation: With simple randomization, we might randomly get imbalances on important characteristics (especially in small samples). Can we do better?

Simple randomization: Randomly assign all units to treatment or control

Stratified randomization: Divide sample into strata (blocks) based on pre-treatment characteristics, then randomize **within each stratum**

Example: Study of job training program

- Stratify by: education level (high school vs. college) and gender
- Creates 4 strata: HS males, HS females, College males, College females
- Within each stratum, randomly assign 50% to treatment, 50% to control

Key difference: Simple randomization draws from entire population; stratified randomization ensures representation from each subgroup

Why use stratified randomization?

Two main benefits:

Why use stratified randomization?

Two main benefits:

1. Ensures balance on stratification variables

- With simple randomization, treatment and control groups may differ on key characteristics (especially with small samples)
- Stratification **guarantees** balance on the stratification variables
- Example: Exactly 50% of treated are female if you stratify by gender

Why use stratified randomization?

Two main benefits:

1. Ensures balance on stratification variables

- With simple randomization, treatment and control groups may differ on key characteristics (especially with small samples)
- Stratification **guarantees** balance on the stratification variables
- Example: Exactly 50% of treated are female if you stratify by gender

2. Increases precision (lowers standard errors)

- If stratification variables predict outcomes, controlling for strata reduces residual variance
- Recall: $V(\hat{\beta}) = \sigma^2(X'X)^{-1}$ where σ^2 is residual variance
- Lower residual variance $\Rightarrow$ smaller standard errors $\Rightarrow$ more statistical power

Implementing stratified randomization

How to choose stratification variables?

- Pick variables that are:
  - Strong predictors of the outcome (increases precision)
  - Measured before randomization (ensures exogeneity)
  - Create a manageable number of strata (rule of thumb: at least 4-6 observations per stratum-treatment combination)

- Common choices: baseline outcome, gender, age groups, geographic location

Implementing stratified randomization

How to choose stratification variables?

- Pick variables that are:
  - Strong predictors of the outcome (increases precision)
  - Measured before randomization (ensures exogeneity)
  - Create a manageable number of strata (rule of thumb: at least 4-6 observations per stratum-treatment combination)
- Common choices: baseline outcome, gender, age groups, geographic location

Specification with stratified randomization:

Include **stratum fixed effects** in your regression:

$$Y_i = \alpha + \beta D_i + \sum_{s=1}^{S} \gamma_s \mathbb{1}[\text{Stratum}_i = s] + u_i \tag{50}$$

- This accounts for how randomization was done
- Improves precision (even though β estimate is similar without FE)

3. Imperfect compliance

- In many cases, people cannot be compelled to participate into the experiment
- Within the experiment, people cannot be obliged to take the treatment → reintroduces selectivity

How to solve these problems?

Example: Krueger class size paper

1. Approx. 10% have changed class type during the experiment upon teacher request (behavioral problems) or parents pressure
2. Some children have changed school or moved (“attrition”): can be correlated both to class type and student characteristics (weak or good)

The design is now an **encouragement design**

Simpler to implement, more acceptable, often no choice

Comes at a cost in precision

Introducing endogenous changes in class composition

- Initial random assignment: $D_i = 0/1$
- Class in which students are actually observed: $T_i = 0/1$

Notations:

$$p_0 = P(T_i = 1 | D_i = 0) \tag{51}$$
$$p_1 = P(T_i = 1 | D_i = 1) \tag{52}$$

What do we observe?

Introducing endogenous changes in class composition

- Initial random assignment: $D_i = 0/1$
- Class in which students are actually observed: $T_i = 0/1$

Notations:

$$p_0 = P(T_i = 1 | D_i = 0) \tag{51}$$
$$p_1 = P(T_i = 1 | D_i = 1) \tag{52}$$

What do we observe?

We observe D_i, T_i, p_0, p_1, and $Y_i(0)$ if $T_i = 0$ and $Y_i(1)$ if $T_i = 1$

Let's simplify notation to see things better:

- Drop the *i*'s
- Let $Y_i(1) := Y_1$; $Y_i(0) := Y_0$

Deriving a Wald Estimator:

$$
\begin{aligned}
\mathbb{E}[Y|D=1] &= \mathbb{E}[Y_0|T=0, D=1]P(T=0|D=1) && (53)\\
&\quad + \mathbb{E}[Y_1|T=1, D=1]P(T=1|D=1) && (54)\\
&= \mathbb{E}[Y_0|T=0, D=1]P(T=0|D=1) && (55)\\
&\quad + \mathbb{E}[Y_0|T=1, D=1]P(T=1|D=1) && (56)\\
&\quad + \mathbb{E}[Y_1 - Y_0|T=1, D=1]P(T=1|D=1) && (57)\\
&= \mathbb{E}[Y_0|D=1] && (58)\\
&\quad + \mathbb{E}[Y_1 - Y_0|T=1, D=1]P(T=1|D=1) && (59)
\end{aligned}
$$

Assuming constant effect model

$$= \mathbb{E}[Y_0] + \mathbb{E}[Y_1 - Y_0]p_1$$

Reminder: Wald estimator

Thus we have

$$\mathbb{E}[Y|D=1] = \mathbb{E}[Y_0] + \mathbb{E}[Y_1 - Y_0]p_1 \tag{60}$$

$$\mathbb{E}[Y|D=0] = \mathbb{E}[Y_0] + \mathbb{E}[Y_1 - Y_0]p_0 \tag{61}$$

Thus

$$\mathbb{E}[Y|D=1] - \mathbb{E}[Y|D=0] = \mathbb{E}[Y_1 - Y_0](p_1 - p_0)$$

and

$$ATE = \mathbb{E}[Y_1] - \mathbb{E}[Y_0] = \frac{\mathbb{E}[Y|D=1] - \mathbb{E}[Y|D=0]}{(p_1 - p_0)}$$

where numerator and denominator are observable

Which populations are comparable?

$D = 0$ *Assigned to large class*	$D = 1$ *Assigned to small class*
Remain: 80% UNTREATED ($T_i = 0$) ———————— Move to small class: 20% TREATED ($T_i = 1$)	TREATED ($T_i = 1$) (100%)

The only populations we can compare “legally” are **ALL** $D_i = 0$ vs **ALL** $D_i = 1$, because these are the only randomly assigned group. The choice of some people to move to small class is endogenous.

Differences in outcome between the 2 random groups?

$D = 0$ *Assigned to large class*	$D = 1$ *Assigned to small class*
1: untreated – 5	5: treated – 17
2: untreated – 5	6: treated – 5
3: untreated – 15	7: treated – 15
	8: treated – 15
—————————	
4: treated – 15	

Intention to treat (ITT)

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

The “intention to treat” is the difference in outcomes comparing those we *intended* to treat or not

- Assigned large class: avg. 10 ($\mathbb{E}[Y_i|D_i = 0]$)
- Assigned small class: avg. 13 ($\mathbb{E}[Y_i|D_i = 1]$)
- $ITT = 13 - 10 = 3$

Questions:

- Is this a causal effect?

Intention to treat (ITT)

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

The “intention to treat” is the difference in outcomes comparing those we *intended* to treat or not

- Assigned large class: avg. 10 ($\mathbb{E}[Y_i|D_i = 0]$)
- Assigned small class: avg. 13 ($\mathbb{E}[Y_i|D_i = 1]$)
- $ITT = 13 - 10 = 3$

Questions:

- Is this a causal effect?
- Is this an interesting thing to estimate?

But what is the effect of the **treatment** (T_i)?

$D = 0$	$D = 1$
Assigned to large class	*Assigned to small class*
1: untreated – 5	5: treated – 17
2: untreated – 5	6: treated – 5
	7: treated – 15
	8: treated – 15
3: treated – 15	
4: treated – 15	

Treatment effect

- Assigned large class: avg. 10 ($\mathbb{E}[Y_i|D_i = 0]$)
- Assigned small class: avg. 13 ($\mathbb{E}[Y_i|D_i = 1]$)
- Assigned large class: 50% treated ($P(T_i = 1|D_i = 0) = p_0$)
- Assigned small class: 100% treated ($P(T_i = 1|D_i = 1) = p_1$)

Y_i is 3 higher in when assigned to treatment (ITT)

by moving only .5 persons from untreated to treated ($p_1 - p_0$)

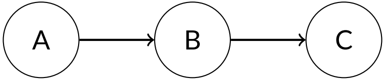
So by how much is the treatment increasing each person's mark?

Treatment effect

$$\begin{aligned} \mathbb{E}[Y_1] - \mathbb{E}[Y_0] &= \frac{\mathbb{E}[Y|D=1] - \mathbb{E}[Y|D=0]}{(p_1 - p_0)} && (62) \\ &= 3/.5 && (63) \\ &= 3 \times 2 = 6 && (64) \end{aligned}$$

Warning: Actually I'm sneaking in some assumptions here, this is actually the *Local Average Treatment Effect* which may not be the same as the *Average Treatment Effect* (more next week!)

Aside: terminology



- **“Reduced form”** ($A \rightarrow C$): we ignore the mechanism and just look at the overall effect on the final outcome
- **“First stage”** ($A \rightarrow B$): we focus on the impact of some intervention on something earlier in the causal chain

This terminology comes up particularly when looking at **instrumental variables** and when trying to understand **mechanisms**

Outline

1. Potential outcomes framework

2. Selection bias

3. Controlled experiments

4. Clustered randomization

5. Statistical power

6. Conclusion

Beyond individual randomization

- So far: randomization at the **individual level**
- Sometimes randomizing individuals is:
  - **Impractical**: Hard to treat some students in a school but not others
  - **Unethical**: Perceived as unfair within communities
  - **Contaminated**: Treatment spills over between individuals

Beyond individual randomization

- So far: randomization at the **individual level**
- Sometimes randomizing individuals is:
  - **Impractical**: Hard to treat some students in a school but not others
  - **Unethical**: Perceived as unfair within communities
  - **Contaminated**: Treatment spills over between individuals
- **Solution**: Randomize at a higher level - **cluster randomization**
- Treat entire groups (clusters) as units: schools, villages, clinics, firms

Examples of clustered randomization

- **Education**: Randomize schools (not students)
  - Teacher training programs
  - School infrastructure improvements
- **Health**: Randomize clinics or villages
  - Deworming programs (Miguel & Kremer 2004)
  - Community health worker programs
- **Development**: Randomize villages or districts
  - Microfinance expansion
  - Infrastructure projects (roads, electricity)

Examples of clustered randomization

- **Education**: Randomize schools (not students)
  - Teacher training programs
  - School infrastructure improvements

- **Health**: Randomize clinics or villages
  - Deworming programs (Miguel & Kremer 2004)
  - Community health worker programs

- **Development**: Randomize villages or districts
  - Microfinance expansion
  - Infrastructure projects (roads, electricity)

Key insight: The unit of randomization $\neq$ unit of analysis

Spillovers and SUTVA violations

Recall our SUTVA assumption: Y_i depends only on own treatment D_i

With spillovers, potential outcomes become:

$$Y_i(D_i, \mathbf{D}_{-i}) \tag{65}$$

where $\mathbf{D}_{-i}$ is the treatment status of others

Spillovers and SUTVA violations

Recall our SUTVA assumption: Y_i depends only on own treatment D_i

With spillovers, potential outcomes become:

$$Y_i(D_i, \mathbf{D}_{-i}) \tag{65}$$

where $\mathbf{D}_{-i}$ is the treatment status of others

Example: Deworming

- Direct effect: Health benefits to treated children
- **Spillover**: Reduced disease transmission to untreated children
- Individual randomization would **underestimate** total effect

Spillovers and SUTVA violations

Recall our SUTVA assumption: Y_i depends only on own treatment D_i

With spillovers, potential outcomes become:

$$Y_i(D_i, \mathbf{D}_{-i}) \tag{65}$$

where $\mathbf{D}_{-i}$ is the treatment status of others

Example: Deworming

- Direct effect: Health benefits to treated children
- **Spillover**: Reduced disease transmission to untreated children
- Individual randomization would **underestimate** total effect

Cluster randomization partially solves this:

- Captures within-cluster spillovers
- But still misses cross-cluster spillovers

Notation for clustered designs

Let $c = 1, \dots, C$ index clusters, $i = 1, \dots, N_c$ index individuals within cluster c

- $D_c \in \{0, 1\}$: treatment status of cluster c
- Y_{ic}: outcome for individual i in cluster c
- All individuals in cluster c receive same treatment

Notation for clustered designs

Let $c = 1, \ldots, C$ index clusters, $i = 1, \ldots, N_c$ index individuals within cluster c

- $D_c \in \{0, 1\}$: treatment status of cluster c
- Y_{ic}: outcome for individual i in cluster c
- All individuals in cluster c receive same treatment

Potential outcomes:

$$Y_{ic}(1) = \text{outcome if cluster } c \text{ is treated} \tag{66}$$

$$Y_{ic}(0) = \text{outcome if cluster } c \text{ is not treated} \tag{67}$$

Notation for clustered designs

Let $c = 1, ..., C$ index clusters, $i = 1, ..., N_c$ index individuals within cluster c

- $D_c \in \{0, 1\}$: treatment status of cluster c
- Y_{ic}: outcome for individual i in cluster c
- All individuals in cluster c receive same treatment

Potential outcomes:

$$Y_{ic}(1) = \text{outcome if cluster } c \text{ is treated} \tag{66}$$

$$Y_{ic}(0) = \text{outcome if cluster } c \text{ is not treated} \tag{67}$$

Observed outcome:

$$Y_{ic} = D_c Y_{ic}(1) + (1 - D_c) Y_{ic}(0) \tag{68}$$

Note: Everyone in the cluster has the same D_c!

Estimation with clustering

Simple comparison of means still works:

$$\widehat{ATE} = \bar{Y}_{treated} - \bar{Y}_{control} \tag{69}$$

Estimation with clustering

Simple comparison of means still works:

$$\widehat{ATE} = \bar{Y}_{treated} - \bar{Y}_{control} \tag{69}$$

BUT: Standard errors must account for clustering!

Why? Outcomes within clusters are correlated:

- Students in same school face same teachers, facilities
- Villagers share local economic conditions
- This reduces **effective sample size**

Estimation with clustering

Simple comparison of means still works:

$$\widehat{ATE} = \bar{Y}_{treated} - \bar{Y}_{control} \tag{69}$$

BUT: Standard errors must account for clustering!

Why? Outcomes within clusters are correlated:

- Students in same school face same teachers, facilities
- Villagers share local economic conditions
- This reduces **effective sample size**

Intuition: 1000 students in 10 schools provides **less information** than 1000 randomly selected students

Estimation with clustering

Simple comparison of means still works:

$$\widehat{ATE} = \bar{Y}_{treated} - \bar{Y}_{control} \tag{69}$$

BUT: Standard errors must account for clustering!

Why? Outcomes within clusters are correlated:

- Students in same school face same teachers, facilities
- Villagers share local economic conditions
- This reduces **effective sample size**

Intuition: 1000 students in 10 schools provides **less information** than 1000 randomly selected students

⇒ Use **cluster-robust standard errors** that account for within-cluster correlation

Design trade-offs: Number vs size of clusters

For a fixed total sample size N, how to allocate across clusters?

Key parameter: Intra-cluster correlation (ICC) $= \rho$

- ρ = correlation between outcomes of individuals in same cluster
- $\rho = 0$: no clustering, like individual randomization
- $\rho = 1$: everyone in cluster identical

Design trade-offs: Number vs size of clusters

For a fixed total sample size *N*, how to allocate across clusters?

Key parameter: Intra-cluster correlation (ICC) $= \rho$

- ρ = correlation between outcomes of individuals in same cluster
- $\rho = 0$: no clustering, like individual randomization
- $\rho = 1$: everyone in cluster identical

Design effect (variance inflation):

$$DE = 1 + (n - 1)\rho \tag{70}$$

where *n* = average cluster size

Design trade-offs: Number vs size of clusters

For a fixed total sample size N, how to allocate across clusters?

Key parameter: Intra-cluster correlation (ICC) $= \rho$

- ρ = correlation between outcomes of individuals in same cluster
- $\rho = 0$: no clustering, like individual randomization
- $\rho = 1$: everyone in cluster identical

Design effect (variance inflation):

$$DE = 1 + (n - 1)\rho \tag{70}$$

where n = average cluster size

Implications:

- More clusters $>$ bigger clusters (for statistical power)
- If $\rho = 0.05$ and $n = 20$: need $\approx 2\times$ the sample size!
- Rules of thumb: Need at least 20-30 clusters for reliable inference

Advantages and disadvantages

Advantages of cluster randomization:

- Captures within-cluster spillovers
- Administratively simpler
- More acceptable to communities
- Can study cluster-level interventions

Advantages and disadvantages

Advantages of cluster randomization:

- Captures within-cluster spillovers
- Administratively simpler
- More acceptable to communities
- Can study cluster-level interventions

Disadvantages:

- Lower statistical power
- Requires more clusters for balance
- Still misses cross-cluster spillovers
- Harder to study heterogeneous effects

Advantages and disadvantages

Advantages of cluster randomization:

- Captures within-cluster spillovers
- Administratively simpler
- More acceptable to communities
- Can study cluster-level interventions

Disadvantages:

- Lower statistical power
- Requires more clusters for balance
- Still misses cross-cluster spillovers
- Harder to study heterogeneous effects

Bottom line: Use clustered randomization when spillovers matter or individual randomization is infeasible

Wald and instrumental variable

Implementation on simultaneous equation model:

$$T = \pi_0 + \pi_1 D + \varepsilon \tag{71}$$

$$Y = \beta_0 + \alpha_1 T + u \tag{72}$$

D exogenous, but T endogenous (it is a choice)

Notice

$$\pi_0 = p_0 \tag{73}$$

$$\pi_1 = p_1 - p_0 \tag{74}$$

"Reduced form" (meaning we ignore the mechanism and just look at the effect of assignment on outcome)

$$Y = \beta_0 + \alpha_1[\pi_0 + \pi_1 D + \varepsilon] + u \tag{75}$$

$$= \beta_0 + \alpha_1 \pi_0 + \alpha_1 \pi_1 D + u' \tag{76}$$

Reduced form:

$$Y = \beta_0 + \alpha_1\pi_0 + \alpha_1\pi_1 D + u'$$

- Regressing Y on D estimates $\alpha_1\pi_1$ (no problem because D is exogenous)
- $\alpha_1\pi_1 = \mathbb{E}[Y|D_i = 1] - \mathbb{E}[Y|D = 0]$
- ITT is the reduced form

First stage:

$$T = \pi_0 + \pi_1 D + \varepsilon \tag{77}$$

- Regressing T on D_i estimates π_1 (no problem because D_i exogenous)
- $\pi_1 = p_1 - p_0$

From 2 equations where D_i is exogenous, we can identify

$$\frac{\alpha_1 \pi_1}{\pi_1}$$

α_1 is thus identified by $[\alpha_1 \pi_1 / \pi_1]$

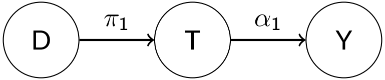
This is exactly the Wald estimator

From 2 equations where D_i is exogenous, we can identify

$$\frac{\alpha_1 \pi_1}{\pi_1}$$

α_1 is thus identified by $[\alpha_1 \pi_1 / \pi_1]$

This is exactly the Wald estimator



Instrumental variable estimator (2SLS)

OLS on the endogenous variable:

$$\hat{T} = \hat{\pi}_0 + \hat{\pi}_1 D$$

OLS on the structural equation, substituting T with its predicted value:

$$Y = \beta_0 + \alpha_1 \hat{T} + u$$

- Deviations from initial random assignment are not an issue because we have a very robust instrumental variable: D (but there can be substantial efficiency losses)
- Naturally, the initial random assignment has to determine strongly the actual treatment (i.e. $\pi_1 \neq 0$ or $p_1 \neq p_0$)

2SLS estimation (T: number of students)

Table VII
OLS and 2SLS Estimates of Effect of Class Size on Achievement
Dependent Variable: Average Percentile Score on SAT

Grade	OLS (1)	2SLS (2)	Sample size (3)
K	-.62	-.71	5,861
	(.14)	(.14)	
1	-.85	-.88	6,452
	(.13)	(.15)	
2	-.59	-.67	5,950
	(.12)	(.14)	
3	-.61	-.81	6,109
	(.13)	(.15)	

Summarising

So we have two main ways of extracting causal estimates from RCTs with imperfect compliance:

- **Intention to Treat**: the reduced-form effect of treatment assignment on the outcome
- **Instrumental Variable / 2SLS / Wald Estimate**: the effect of actually receiving the treatment T on the outcome (but only on compliers – see next week)

Question: Which of these is more useful? When?

The cost of non-compliance

What is the cost of low compliance if we are using Wald Estimate?

$\rightarrow$

The cost of non-compliance

What is the cost of low compliance if we are using Wald Estimate?
$\rightarrow$ Lower precision in our estimates

The cost of non-compliance

What is the cost of low compliance if we are using Wald Estimate?
→ Lower precision in our estimates

With full compliance, estimate β by OLS

$$Y = \alpha + \beta D_i + u$$

Leave as an exercise to show that

$$V(\hat{\beta}) = \frac{1}{\bar{D}(1-\bar{D})}\frac{V(u)}{N}$$

The cost of non-compliance

With non-compliance, it can be shown that

$$V(\hat{\beta}_{IV}) = \frac{1}{\bar{D}(1-\bar{D})}\frac{V(u)}{N}\frac{1}{\pi_1^2}$$

where $\pi_1 = p_1 - p_0$ ie the fraction of assigned that take treatment minus the fraction of non-assigned that take tretement

So the standard error is higher by a factor π_1.

Implication for design

You have population N_0 likely to be treated

You anticipate that a share π_1 will accept the treatment

Two options:

1. Randomize 50% of N_0 and have compliance π_1
2. Ask for volonteers and then randomize

Implication for design

You have population N_0 likely to be treated

You anticipate that a share π_1 will accept the treatment

Two options:

1. Randomize 50% of N_0 and have compliance π_1
2. Ask for volonteers and then randomize

Latter: sample of $\pi_1 N_0$ but full compliance

Ratio of variances:

$$\frac{V_1}{V_2} = \frac{1/\pi_1^2 N_0}{1/\pi_1 N_0} = \frac{1}{\pi_1}$$

Implication for design

Lesson: more precise estimates if you randomize among a (smaller) population of people who are likely to take up the treatment

Intuition:

Implication for design

Lesson: more precise estimates if you randomize among a (smaller) population of people who are likely to take up the treatment

Intuition: With low compliance, you **don't know** who exactly was induced to comply by the treatment (D is a "blurry lever" for T) which reduces precision. But if you randomize among volunteers, you know exactly who is complying.

Outline

1. Potential outcomes framework

2. Selection bias

3. Controlled experiments

4. Clustered randomization

5. Statistical power

6. Conclusion

Designing an experiment

Two main questions when designing an experiment:

1. Who to randomize, how, etc.
2. Sample size (and share treated)

Experiments are an unusual case where you have great control over sample size

The last thing you want: go through the whole burden and have insignificant effects because you have high standard errors

Finite sample and inference

So far, we have always considered the asymptotic values of the estimator

For instance:

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

is the asymptotic value of:

$$\frac{1}{N_1} \sum_{i \in D_i = 1} y_i$$

which, inversely, is the empirical counterpart to $\mathbb{E}[Y_i|D_i = 1]$

This is because we have been interested in **identification** (what we would learn in infinite samples)

Finite sample and inference

Random experiment: T and C are similar for $N = \infty$

In finite samples, T and C always *somewhat* different, e.g. by chance my treatment group has slightly older students than the control group

This **imbalance** could be confounded with treatment effect

Inference is accounting for that:
With finite sample, can I consider that the difference T vs. C is high enough to indicate more than unavoidable imbalance?
Yes, if statistically “significant”

Imbalance is not a source of bias; the standard error is there to account for that

Reminder: significance tests

Estimator $\hat{\beta}$ asymptotically normal with mean β and variance $V(\hat{\beta}) = \sigma_{\beta}^2$

If $\beta = 0$, then, for a risk α (e.g. 5%) we can define $t_{\alpha/2}$ such that:

$$P\left(-t_{\alpha/2} < \frac{\hat{\beta}}{\sigma_{\beta}} < t_{\alpha/2}\right) = 1 - \alpha$$

Thus

$$2\Phi(t_{\alpha/2}) - 1 = 1 - \alpha$$

and we can read $t_{\alpha/2}$ for the normal distribution table

For $\alpha = 0.05$, $\Phi(1.96) = 0.975$

If $|\hat{\beta}/\sigma_{\beta}| > 1.96$, we can reject the null $\beta = 0$

Balance table

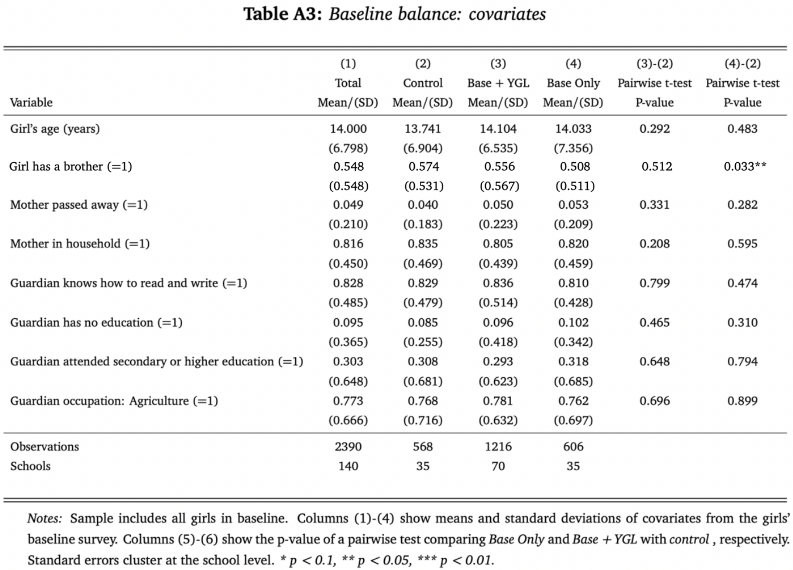


Table A3: *Baseline balance: covariates*

	(1)	(2)	(3)	(4)	(3)-(2)	(4)-(2)
	Total	Control	Base + YGL	Base Only	Pairwise t-test	Pairwise t-test
Variable	Mean/(SD)	Mean/(SD)	Mean/(SD)	Mean/(SD)	P-value	P-value
Girl's age (years)	14.000	13.741	14.104	14.033	0.292	0.483
	(6.798)	(6.904)	(6.535)	(7.356)		
Girl has a brother (=1)	0.548	0.574	0.556	0.508	0.512	0.033**
	(0.548)	(0.531)	(0.567)	(0.511)		
Mother passed away (=1)	0.049	0.040	0.050	0.053	0.331	0.282
	(0.210)	(0.183)	(0.223)	(0.209)		
Mother in household (=1)	0.816	0.835	0.805	0.820	0.208	0.595
	(0.450)	(0.469)	(0.439)	(0.459)		
Guardian knows how to read and write (=1)	0.828	0.829	0.836	0.810	0.799	0.474
	(0.485)	(0.479)	(0.514)	(0.428)		
Guardian has no education (=1)	0.095	0.085	0.096	0.102	0.465	0.310
	(0.365)	(0.255)	(0.418)	(0.342)		
Guardian attended secondary or higher education (=1)	0.303	0.308	0.293	0.318	0.648	0.794
	(0.648)	(0.681)	(0.623)	(0.685)		
Guardian occupation: Agriculture (=1)	0.773	0.768	0.781	0.762	0.696	0.899
	(0.666)	(0.716)	(0.632)	(0.697)		
Observations	2390	568	1216	606		
Schools	140	35	70	35		

Notes: Sample includes all girls in baseline. Columns (1)-(4) show means and standard deviations of covariates from the girls' baseline survey. Columns (5)-(6) show the p-value of a pairwise test comparing *Base Only* and *Base + YGL* with *control* , respectively. Standard errors cluster at the school level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Overall balance

Often, because we are doing **multiple hypothesis tests** we will get a few significant imbalances when looking across multiple outcomes

How can we test for **overall** imbalance?

Overall balance

Often, because we are doing **multiple hypothesis tests** we will get a few significant imbalances when looking across multiple outcomes

How can we test for **overall** imbalance?

1. Run regression:

$$\text{Treated}_i = \beta_0 + \beta_1 \text{Outcome1}_i + \beta_2 \text{Outcome2}_i + \ldots + \beta_K \text{OutcomeK}_i \tag{78}$$

2. Use an F-test (joint test) of $\beta_1 = \beta_2 = \ldots = \beta_K = 0$

The power of the experiment

If the policy has an impact, we want to be able to see it

Unless the effect is very small, we want to reject the null

But if there is a lot of imprecision (large estimator variance), we may fail to do so

Type II error is the probability of a **false negative**, i.e., $\beta > 0$, but we fail to reject the null ($\hat{\beta}/\sigma_{\beta} < 1.96$). In other words, we fail to detect an effect that is really there.

This will happen sometimes, for some samples

Power = $1 - P$(Type II error), i.e. the probability that we detect an effect if there really is one.

The power of the experiment

Usual approach: set an acceptable power (typically 80%), and then:

1. Set a reasonable β that you feel you should be able to "see" (the **minimum detectable effect** you want
2. And figure out the sample size that ensures that power for a true effect β

Computing the power

Let's calculate the power, where $(\hat{\beta}/\sigma_\beta < 1.96)$ and β is random:

$$P\left(\frac{\hat{\beta}}{\sigma_\beta} > t_{\alpha/2}|\beta\right) = \kappa$$

where κ is the power.

$$P\left(\frac{\hat{\beta}-\beta}{\sigma_\beta} > t_{\alpha/2} - \frac{\beta}{\sigma_\beta}|\beta\right) = \kappa$$

$$\Phi\left(\frac{\beta}{\sigma_\beta} - t_{\alpha/2}\right) = \kappa$$

Thus:

$$\frac{\beta}{\sigma_\beta} - t_{\alpha/2} = t_{1-\kappa}$$

Minimum detectable effect

The β that will be "significant" 80% of the time (at 5% level) is such that:

$$\frac{\beta}{\sigma_\beta} - t_{\alpha/2} = t_{1-\kappa}$$

or

$$\beta = (t_{\alpha/2} + t_{1-\kappa})\sigma_\beta$$

with $t_{\alpha/2} = 1.96$ if $\alpha = 0.05$ and $t_{1-\kappa} = 0.84$ if $\kappa = 0.80$

$(t_{\alpha/2} + t_{1-\kappa})\sigma_\beta$ is the **minimum detectable effect** (MDE)

MDE and sample size

Consider the model:

$$y = c + \beta D_i + u$$

Remember that:

$$\sigma_{\beta}^2 = \frac{1}{\bar{D}(1-\bar{D})} \frac{V(u)}{N}$$

Thus:

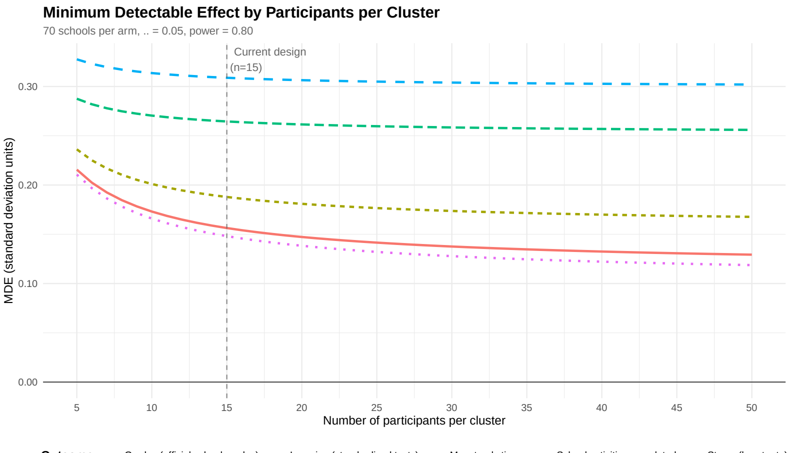
$$\text{MDE} = (t_{\alpha/2} + t_{1-\kappa}) \sqrt{\frac{1}{\bar{D}(1-\bar{D})} \frac{V(u)}{N}}$$

Interpret each of those terms... (think in terms of finite sample imbalance)

How does MDE increase with sample size?

MDE and cluster size

Key insight: With clustered randomization, increasing the number of people surveyed in each cluster doesn't decrease MDE much



With instrumental variables

$$y = c + \beta T + u$$

where treatment T is instrumented by some random assignment D_i

Reminder:

$$V(\hat{\beta}_{IV}) = \frac{1}{\bar{D}(1-\bar{D})} \frac{V(u)}{N} \frac{1}{\pi_1^2}$$

The precision decreases linearly with the (net) take-up

So does the MDE

If take-up is 50%, implies more than doubling sample size.

Outline

1. Potential outcomes framework

2. Selection bias

3. Controlled experiments

4. Clustered randomization

5. Statistical power

6. Conclusion

Summing up

1. **Potential outcomes framework -** this gives us a way of conceptualizing counterfactuals and articulating clearly when we can and cannot make causal claims
2. **Randomized controlled trials** are a way to make causal claims with relatively weak assumptions on the data generating process
3. **Design issues -** we learnt about various design issues that come up in RCTs, e.g., dealing with imbalances, calculating statistical power, dealing with imperfect compliance

Temporary page!

LaTeX was unable to guess the total number of pages correctly. As there was some unprocessed data that should have been added to the final page this extra page has bee added to receive it.

If you rerun the document (without altering it) this surplus page will go away, because LA now knows how many pages to expect for this document.