ISEA Week 9 – Causal Inference, RCT, A/B Testing

Dr. Chris Candelaria, Vanderbilt University

IES Institute of Education Sciences

W AmplifyLearn.AI

UNIVERSITY of WASHINGTON
eScience Institute
ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS

O UNIVERSITY OF OREGON

Agenda

1. Logistics
2. A/B testing vs RCTs
3. RCTs in Education Policy

Logistics

1. Lodging and travel for Hackweek.
2. Mid-session fellows’ feedback and adjustment.
3. ISEA will offer Digital Badge to fellows who will turn in homework, attend 90% of the web sessions, and use tutoring sessions to practice coding, and complete Hackweek.
4. Hackweek recruitment is [here.]() Distribute widely or submit your own project.
5. Video posting.
6. Please use all the resources available to you! Webinar is not the only thing!
    a. Teams Channel for Peer Learning, messaging, and sharing resources.
    b. Study Hall for Peer Learning.
    c. Tutoring sessions for practicing and learning about coding.
    d. Turn-in homeworks to receive feedback.
    e. Extra study materials will be provided through Teams, Canvas, and Footnotes on session slidedeck.

A/B Testing

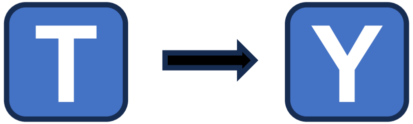
1. Study Design
2. Decide what features to test
3. Recruit participants
4. Determine the length of one round of A/B testing (4 weeks, Kohavvi et al., 2023)
5. Design what data to collect and how: Google Analytics and surveys.
6. Analyze data in rapid cycle: t-test; regression
7. Decide how many rounds
8. Conduct power analysis to find the ideal sample size if possible. In reality, get as many users as possible, because the sample size varies depending on how sensitive a tested feature is, the unit of the analysis (at individual user level or a given task level), and the duration of the testing session.
9. Minimizing contamination

Outcome differences between A/B tests and RCTs: Let’s Brainstorm!

A/B Testing Outcomes

RCT Education Outcomes

Randomized Controlled Trials (RCT): The Problem of Selection Bias



Potential Outcomes Framework: **(Holland, 1986)**

$$Y_i = Y_i(0) + T_i(Y_i(1) - Y_i(0))$$

If $T_i = 1$: $Y_i = Y_i(1)$

$T_i = 0$: $Y_i = Y_i(0)$

Assuming constant treatment effect:

$$Y_i(1) = Y_i(0) + \tau$$

Average Treatment Effect (ATE):

$$E[Y_i(1) - Y_i(0)] = \tau$$

Use difference in averages to estimate the ATE?

$$E[Y_i|T_i = 1] - E[Y_i|T_i = 0]$$

$$= E[Y_i(1)|T_i = 1] - E[Y_i(0)|T_i = 0]$$

$$= E[Y_i(0) + \tau|T_i = 1] - E[Y_i(0)|T_i = 0]$$

$$= \underset{\substack{\uparrow \\ \text{ATE}}}{\tau} + \underbrace{E[Y_i(0)|T_i = 1] - E[Y_i(0)|T_i = 0]}_{\text{Selection Bias!}}$$

RCTs remove selection bias. Why?

RCTs: A Population Regression Framework

- **If** $Y_i = \beta_0 + \beta_1 T_i + u_i$ **and** $E[u_i|T_i] = 0$
- **Then:**

$$E[Y_i|T_i = 1] \\ = E[\beta_o + \beta_1(1) + u_i|T_i = 1] \\ = \beta_0 + \beta_1 + E[u_i|T_i = 1] \\ = \beta_0 + \beta_1$$

$$E[Y_i|T_i = 0] \\ = E[\beta_o + \beta_1(0) + u_i|T_i = 0] \\ = \beta_0 + E[u_i|T_i = 0] \\ = \beta_0$$

$$\beta_1 = E[Y_i|T_i = 1] - E[Y_i|T_i = 0]$$

RCTs: OLS Estimator

- > **If** $Y_i = \beta_0 + \beta_1 T_i + u_i$ **and** $E[u_i|T_i] = 0$
- > **Then:**

$$\beta_1 = \boxed{E[Y_i|T_i = 1]} - \boxed{E[Y_i|T_i = 0]}$$

OLS Estimator: $\widehat{\beta_1} = E[Y_i\widehat{|T_i =} 1] - E[Y_i\widehat{|T_i =} 0]$

$$= \quad \overline{Y_i}|T_i = 1 - \overline{Y_i}|T_i = 0$$

Unbiased Estimator: $E\left[\widehat{\beta_1}\right] = \beta_1$ **We identify the Average Treatment Effect (ATE)**

IES Institute of Education Sciences W AmplifyLearn.AI UNIVERSITY of WASHINGTON eScience Institute ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS O UNIVERSITY OF OREGON

RCTs: Inference

- Before implementing an RCT or A/B test, need to ensure you have sufficient power to conduct hypothesis tests:
  - *Power*: Probability of rejecting the null hypothesis, given that the null hypothesis is false
  - One way to increase power: increase the sample size
  - Typical power threshold: 80 percent

- **Resource**: PowerUp!
  - https://www.causalevaluation.org/power-analysis.html

RCTs in Education Policy Settings

Before we think about methodology, let's consider some potential blemishes of the so-called gold standard:

> Ethical issues

> Compliance/Fidelity to treatment

> Attrition

> Experiment bias

> External validity

Mechanisms
?

RCT Example: FAFSA Article (Bettinger et al., 2012)

- **In education policy, there has been a push to increase college going rates**
  - Example: The Tennessee Higher Education Commission set goals to increase the college-going rate for the class of 2023 to at least 60 percent
- **However, there are substantive impediments that affect college access**
  - Process of applying for both financial aid and college is difficult
  - Misinformation about the true cost of college
- **Education research: Could a simplified application process significantly improve college going rates?**
  - Goal: Reduce asymmetric information to help students (and parents)

Bettinger et al. (2012): RCT Design

Authors examine whether an intervention of information or direct assistance in filling out financial aid could improve college going rates

> ***H&R Block experiment***

- Approximately 17,000 individuals
- Individuals came from households earning less than $45,000 a year with at least one household member between the ages of 15 and 30 without a bachelor’s degree
- Participation Gift: $20
- Randomization to treatment based on social security number

> ***Three treatment arms/groups***

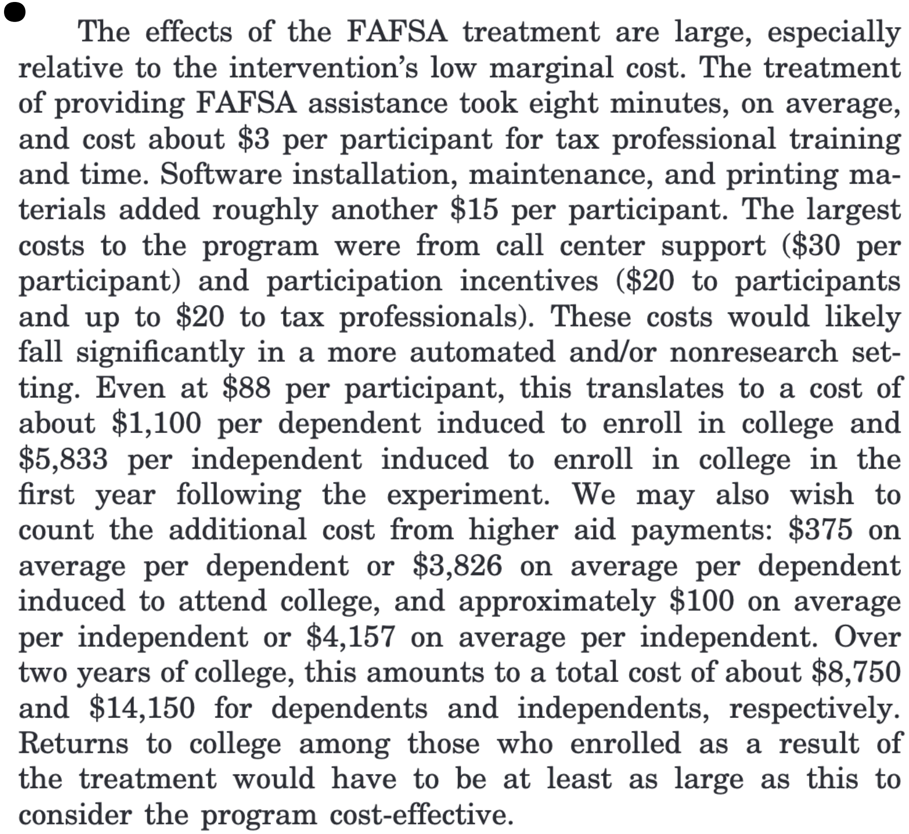
1. (A) Personal assistance in filling out financial aid form and filing it; (B) Information
2. Information only: Potential financial aid amounts and tuition estimates for local colleges
3. Control group: Information on the importance of college and financial aid brochure

Bettinger et al. (2012): Results

- Relative to the control group, 17-year-old high school seniors who received the FAFSA intervention more likely to:
  - *Submit the FAFSA*: **39% increase** (56 vs. 40 percent)
  - *Attend college*: **7 percentage point increase** (34 vs. 27 percent)
  - *Enroll in college for two consecutive years*: **8 percentage point increase** (36 vs. 28 percent)
- No significant differences between information-only and control groups

Bettinger et al. (2012): Costs/Benefits(?)

- $88/participant in the research setting
- Total estimated cost for dependent over 2 years of college: $8,750, on average
- Are the returns to college at least as large as this?



The effects of the FAFSA treatment are large, especially relative to the intervention's low marginal cost. The treatment of providing FAFSA assistance took eight minutes, on average, and cost about $3 per participant for tax professional training and time. Software installation, maintenance, and printing materials added roughly another $15 per participant. The largest costs to the program were from call center support ($30 per participant) and participation incentives ($20 to participants and up to $20 to tax professionals). These costs would likely fall significantly in a more automated and/or nonresearch setting. Even at $88 per participant, this translates to a cost of about $1,100 per dependent induced to enroll in college and $5,833 per independent induced to enroll in college in the first year following the experiment. We may also wish to count the additional cost from higher aid payments: $375 on average per dependent or $3,826 on average per dependent induced to attend college, and approximately $100 on average per independent or $4,157 on average per independent. Over two years of college, this amounts to a total cost of about $8,750 and $14,150 for dependents and independents, respectively. Returns to college among those who enrolled as a result of the treatment would have to be at least as large as this to consider the program cost-effective.

IES Institute of Education Sciences · W · AmplifyLearn.AI · UNIVERSITY of WASHINGTON eScience Institute ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS · UNIVERSITY OF OREGON

Setting Standards: What Works Clearinghouse (WWC)

To assess the strength of an RCT in education policy, WWC follows 5 steps

1. Review outcome measures and check for confounding factors
2. Assess the assignment to treatment conditions
3. Assess compositional change
4. Meet a baseline equivalence standard

IES Institute of Education Sciences W AmplifyLearn.AI UNIVERSITY of WASHINGTON eScience Institute ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS O UNIVERSITY OF OREGON

Step 1: Review outcome measures and check for confounding factors

Outcome Measure Standards

- Standard 1: Face Validity
  - Does the outcome measure what it claims to measure?
- Standard 2: Reliability (Concept from classical test theory)
  - Does the measure yield similar results/scores across different administrations?
- Standard 3: Not over-aligned
  - Does the outcome measure privilege one randomization group over the other?
  - Is the outcome tailored to the treatment condition?
- Standard 4: Consistent data collection procedures

Step 1: Review outcome measures and check for confounding factors

Confounding factors

- According to WWC, A confounding factor has the following characteristics:
  - It is observed
  - It aligns completely with only one study condition
  - It is not part of the intervention the study is testing

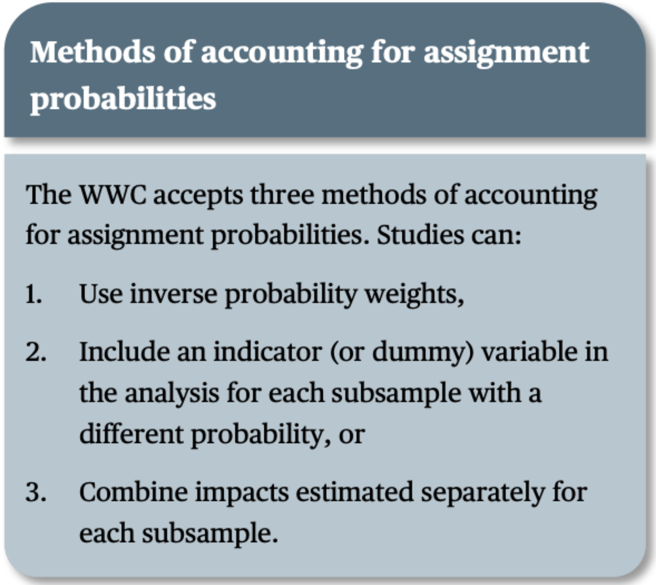
  Source: WWC Manual, Version 5.0 (p. 29)

- Examples?

Step 2: Assess the assignment to treatment conditions

Random assignment is properly done if one of the following two conditions are met:

1. Unit assignment to treatment/control is entirely random (e.g., random number generator)
2. If unit assignment to treatment/control not entirely random, there must be a nonzero probability of being assigned to the conditions



Methods of accounting for assignment probabilities

The WWC accepts three methods of accounting for assignment probabilities. Studies can:

1. Use inverse probability weights,
2. Include an indicator (or dummy) variable in the analysis for each subsample with a different probability, or
3. Combine impacts estimated separately for each subsample.

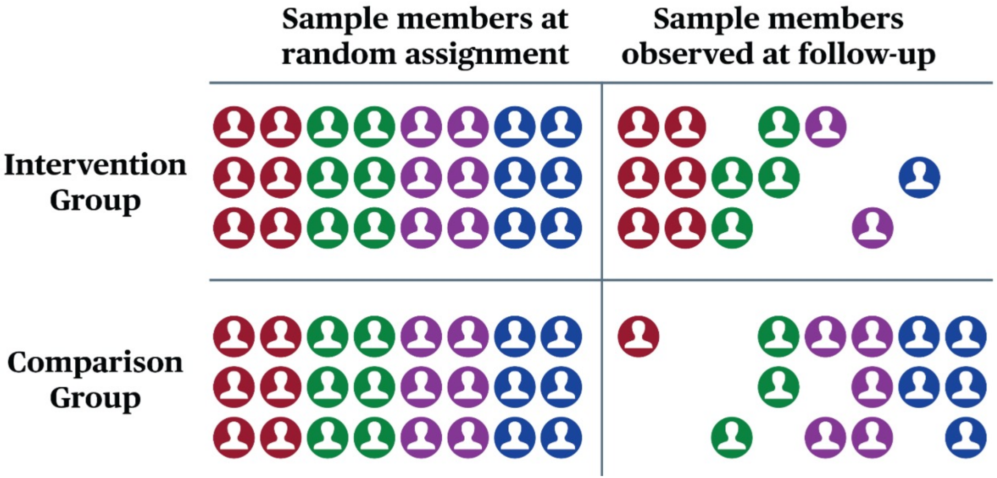
Source: WWC Manual, Version 5.0 (p. 32)

Step 3: Assess compositional change

- Key idea: To what extent might sample attrition affect estimated results of the intervention or treatment?

Two types of attrition:

1. Overall attrition
2. Differential attrition


Sample members at random assignment
Sample members observed at follow-up
Intervention Group
Comparison Group

Source: Figure 6. WWC Procedures and Standards Handbook, Version 5.0

Step 4: Meet a baseline equivalence standard

- Main idea: There should be no differences, on average, between treatment and control groups
  - Observed characteristics and unobserved characteristics

- Strategy: Collect baseline data and test whether groups are balanced on observable characteristics

- But what about unobservable characteristics?

Time to code!

- Access the Google Colab site for our coding session

Assignment Options

- Design an educational RCT:
  - Consider how you would design your study to satisfy the four requirements from the WWC
  - Use PowerUp! to determine the sample size needed for a given effect size
- Work through the school district hypothetical exercise posted on Canvas
- Experiment with the Google Collab Code: Alter the simulation parameters to create your own RCT data set for analysis