Methods for answering causal questions in education research

Ben Castleman
castleman@virginia.edu
My background
Causal analysis defined

A set of methods to isolate the unique effects of the programs we’re interested in from all the other factors that influence the outcomes we care about
Motivation for causal analysis

- School leaders and policy-makers face increasingly difficult decisions about where to invest limited resources.

  Headlines within 24 hours of when I made this slide:
  “Governor Quinn Proposes Education Cuts in Budget”
  “More Cuts Will Come to Prince Williams’ School Budget”
  “Meeting Tonight Will Outline $2M in School Cuts”

- Consider the investment dilemma of a HS principal:
  - More math prep for students behind grade level?
  - More social workers for students with challenges at home?
  - More college counseling for high school seniors?
Causal inference for the data analyst

• As analysts, we’d like to give school leaders definitive info about the impact of various investments

• But how do we separate out the impacts of the programs from the other factors that impact relevant outcomes?

Quality teachers

Other factors impacting learning

Students’ learning
Does watching Elmo lead to children reading earlier?

Are the reading improvements we observe due to Elmo, or is there something about parent involvement that relates to both Elmo watching and reading?
Causal inference to the rescue!

Causal inference methods provide us with a set of strategies to untie the knot and isolate the unique impact of programs we’re analyzing on outcomes that matter.
A sci-fi approach to program evaluation

The causal impact of good teachers: Outcomes for 1\textsuperscript{st} Ben minus the Outcomes for 2\textsuperscript{nd} Ben?
Program evaluation in the real world

• Unfortunately, we haven’t yet figured out how to travel back in time, so we can’t estimate individual treatment effects…

• But what we can look at average treatment effects for a group that are assigned to an intervention (e.g. watching Elmo) compared to a group that does not get an intervention

• All causal methods create or try to find an existing source of random variation in why some kids to participate in a program and others don’t
Why random matters

Random variation means that the group that gets an intervention is equivalent to the group that doesn’t (with some fluctuation)

<table>
<thead>
<tr>
<th>Percent of children…</th>
<th>Assigned to watch Elmo</th>
<th>Assigned to a control group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>56%</td>
<td>56%</td>
</tr>
<tr>
<td>Black</td>
<td>33%</td>
<td>32%</td>
</tr>
<tr>
<td>Low-income</td>
<td>79%</td>
<td>78%</td>
</tr>
<tr>
<td>Reading test score</td>
<td>84</td>
<td>78</td>
</tr>
</tbody>
</table>
A continuum of causal inference

Instrumental Variables Estimation

Randomized Trials:

Regression Discontinuity: Capitalize on an arbitrary cut-off in an index that determines whether students are eligible for a program

Descriptive Analyses

Difference-in-differences: Capitalize on a policy shock that led some students to get a program that others didn’t get
Why are RCTs the gold standard?

1. All other evaluation methods rely on finding a source of random variation in why some people get a program and others don’t—but this is hard!

2. RCTs *force* groups to be equivalent in every respect except who gets the program (or the offer of the program)

3. We then can attribute any downstream differences between groups to the program

4. The results are easy and intuitive to explain
Are RCTs inherently expensive?

• Federal governments conduct high profile and high cost RCTs:
  – Head Start (US gov’t-funded early childcare)
  – Literacy development programs
  – Teacher professional development programs
  – Upward Bound (US gov’t-funded college prep)

• But increasingly access to administrative data helps drive down the cost of RCTs
A conceptual model for thinking about intervention and RCT costs

<table>
<thead>
<tr>
<th>Cost of intervention</th>
<th>Cost of evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Most of my work</td>
<td>Increasingly</td>
</tr>
<tr>
<td>(easy to sell)</td>
<td>popular in US</td>
</tr>
</tbody>
</table>
Randomized Trials - Six basic steps

1. We identify a program we want to evaluate
2. We identify students who are eligible for the program or we recruit students to participate (could also be teachers, schools, etc.)
3. We do a computerized coin flip to decide who gets offered the treatment
4. We confirm the groups are equivalent on baseline characteristics we can observe
5. We implement the program
6. We compare mean outcomes between the treatment and control group
Randomized Trials – An illustrative example

Percentages indicate the share of college-intending students that do not enroll anywhere in the fall following high school graduation.
Step #1: Identify a program to evaluate

• **Strategy:** Hire school counselors to support students during the summer months
Step #2: Identify eligible students

• Eligible sample: Students who had been accepted to and planned to attend college at the end of high school
Step #3: Assign some students to counseling

- Assignment options:
  - First come/first served
  - Student characteristics
  - Teacher nomination
  - Others
  - Coin flip
### Step #4: Confirm groups are equivalent

<table>
<thead>
<tr>
<th>Percent of HS graduates...</th>
<th>Assigned to summer help</th>
<th>Assigned to a control group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>56%</td>
<td>56%</td>
</tr>
<tr>
<td>Black</td>
<td>33%</td>
<td>32%</td>
</tr>
<tr>
<td>Low-income</td>
<td>79%</td>
<td>78%</td>
</tr>
</tbody>
</table>
Step #5: Implement the program
Step #6: Compare mean outcomes between the treatment and control group

Impact of summer outreach on whether students enrolled in college

- Summer help: 59%
- Control: 45%
Extensions to the basic RCT model

1. Examine whether the treatment impacts student sub-groups differentially (e.g. are girls more responsive to additional counseling than boys?)

- **Constant treatment effects**

- **Varying treatment effects**
Extensions to the basic RCT model

2. Evaluate multiple treatment arms or vary the number of students assigned to treatment and control (need to consider sample size)

3. Explore mechanisms that led to program impacts – important to keep detailed participation and interaction data

4. Supplement RCT with qualitative research
Potential limitations in RCTs

1. Too much statistical testing
   – We can conclude randomization failed when it didn’t!
   – We can conclude there are impacts when there aren’t!
   – Solution: Pre-specify analytic plans

2. Too small a sample
   – We can conclude there’s no impact when there is!
   – Solution: Make sure you have a large enough sample to detect reasonable sized impacts
Potential limitations in RCTs

3. Contamination and substitution
   – Members of the treatment “give” the intervention to members of the control group
   – Members of the control group seek out very similar services
   – Solutions: (1) Randomize at a higher level (student vs. school); (2) monitor what the control group does

4. Attrition
   – We only observe outcomes for a subset of the treatment and control groups. These groups may be different, which undermines our randomization
   – Solution: Use administrative data wherever possible
5. Generalizability

- Will programs that work in one context (urban areas) work in all contexts?

- **Solutions:** Replicate studies in different settings

6. Fidelity

- How much variation is there in how a treatment is implemented from one setting to the next? Impacts how we interpret results

- **Solution:** Train and monitor implementers
Practical advice for conducting RCTS
When are RCTs most feasible?

1. When there are more eligible students for a program than available slots

Prior years:

- Students apply from end of junior year until program is full
- Bottom Line confirms academic and financial eligibility
- Bottom Line accepts students until capacity is reached (half of all applicants)

During RCT:

- Students apply by specific deadline
- Bottom Line confirms academic and financial eligibility
- We randomly assign eligible students to the Access program or to a control group
When are RCTs most feasible?

2. When there is a promising innovation in education for which we have little evidence—especially if it’s unobtrusive.
When are RCTs most feasible?

3. When everyone gets something, but what people get varies
   - Important to ensure sufficient contrast

4. When funders prioritize RCTs
Bringing schools on board with RCTs

1. Frame RCTs as a short-term investment for a long-term gain

2. Emphasize randomization as an equitable strategy for deciding who gets the intervention (particularly when resources are constrained)

3. Invest time (as much as it takes) with front-line staff

4. Stagger when people get the treatment
Thanks!

Ben Castleman

castleman@virginia.edu

@BenCastleman
Supplementary slides
Quasi-experimental methods
Difference-in-Differences: Conceptual Overview

• There are many instances in which a policy is implemented in one school (or district or state) but not in other schools
  – e.g. a school adopts extended learning time (ELT)

• We may observe changes in relevant outcomes between the cohorts within the school that got the program and the preceding cohorts that did not get the program
  – e.g. Test scores go up for the cohort that got ELT relative to the preceding cohort that did not have ELT

• But some of these changes in outcomes may have been due to other factors changing over time (e.g. state education policies)

• Diff-in-diffs provides a way to estimate changes in the outcome due the program, less any changes that would have occurred over time
Difference-in-Differences: Conceptual Overview

How much of this change is due to ELT, versus other factors changing over time?

How might we subtract out changes in the outcome due to other factors?

We find a comparison school that is very similar to the school that got ELT, and assume that its change in outcomes captures what the ELT school's change in outcomes would have been in the absence of ELT.

The program impact is then the change in outcomes in the ELT school minus the change in outcomes in the comparison school.
Difference-in-Differences: Key Assumptions

1. The outcome trends in the comparison school capture what the trends in achievement *would have been* in the treated school in the absence of the intervention.

2. There were not other substantial changes happening in the treated school concurrent with the policy we are studying.

3. The change was sufficiently abrupt and unanticipated that we can attribute changes in the outcome to the program we are studying, rather than to compositional changes in the student population, changing behaviors by teachers, etc.
Difference-in-Differences: Issues to consider: Parallel trends?

To address this, we should choose a comparison school (or schools) that had similar historical trends in the outcome as the treated school.
Difference-in-Differences:
Issues to consider: Other changes?

To address this, we should try to understand as much as possible the context where the policy was introduced.
Difference-in-Differences: Issues to consider: Anticipation?

What if the policy change was advertised well in advance of its implementation?

Is the change in outcomes due to ELT, or to a changing student population?

Diff-in-diffs methods are most effective when policy changes are abrupt, substantial, and unanticipated.
Regression Discontinuity: Conceptual Overview

• Often in education, eligibility for a program is determined by where individuals fall on a particular index
  – Need-based scholarships are awarded by whether students fall below a certain income level
  – Kindergarten start dates are determined by whether students turn 5 before a certain date
  – Passage of the 10th grade MCAS is determined by whether students score above a point threshold

• If these cut-offs are arbitrary and unanticipated, they determine whether two groups of students who are otherwise very similar receive a program or not
Regression Discontinuity: Conceptual Overview

The key assumption is that students on either side of a cut-off differ only in that one group of students is eligible for aid and the other is not.
Regression Discontinuity: Additional Assumptions

1. The index cut-off is as good as arbitrary and is only related to the outcome through the particular policy being studied
   - If the cut-off for aid eligibility was the same as the cut-off for food stamps, we could not isolate the impact of aid eligibility

2. The index cut-off should either be unknown to students OR impossible for them to manipulate
   - If students know the aid eligibility cut-off AND can manipulate their income to be just below the cut-off, we shouldn’t expect students on either side of the cut-off to be equivalent
Regression Discontinuity: Implementation

• Authors should be able to convey an entire RD analysis in four pictures
  1. A picture that shows the probability of being eligible for the policy on either side of the cut-off
  2. A picture that shows the density of students on either side of the cut-off
  3. A picture that shows continuity of baseline covariates on either side of the cut-off
  4. A picture that shows a discontinuity in the outcome right at the cut-off

• Results from formal regression analyses should simply reinforce what is conveyed in these pictures
RD Picture #1: Probability of passing a math exam

- All students pass
- All students fail
RD Picture #2:
Density of Observations around the cut-off

Why is the left-hand side good and the right-hand side bad?
RD Picture #3: Continuity of covariates around the cut-off

Why is this a plot we’d like to see? What kind of plot would concern us?
RD Picture #4:
Discontinuity in the outcome at the cut-off

What does this discontinuity imply about the impact of passing a test?
Regression Discontinuity: A limitation

3. About whom can we draw inferences?

RD provides a local average treatment effect: the impact of the policy for students close to the cut-off.

Our inferences apply to people right at the cut-off.

We don’t learn about people far from the cut-off, because we cannot assume they are similar to those right around the cut-off.