FIRST IN THE WORLD (FITW) PROJECT DIRECTORS MEETING:

IMPLEMENTING A SUCCESSFUL QUASI-EXPERIMENTAL DESIGN

MARC MOSS, ABT ASSOCIATES

April 4-5th, 2016
**PLAN FOR THE SESSION**

- Designing a strong evaluation
  - Randomized controlled trial (RCT) brief review

- Quasi-experiment (QED), a viable alternative
  - Forming a valid comparison group
    - Common Pitfalls
    - Attendee discussion
  - Establishing baseline equivalence
    - Attendee discussion
    - How QEDs are reviewed by WWC

- Questions
DESIGNING AND IMPLEMENTING A STRONG EVALUATION

- Goal of a rigorous evaluation is to produce the best evidence of the effectiveness of an intervention

- In theory, evaluations want to know outcomes for a group that received the intervention versus what would have happened to the same group in the absence of the intervention ("counterfactual")

- Counterfactual outcomes are not observed so they have to be estimated.
Plan for the Session

• Designing a strong evaluation
  • Randomized controlled trial (RCT) brief review

• Quasi-experiment (QED), a viable alternative
  • Forming a valid comparison group
    • Common Pitfalls
    • Attendee discussion
  • Establishing baseline equivalence
    • Attendee discussion
    • How QEDs are reviewed by WWC

• Questions
RCTs: **GOLD STANDARD IN THE FIELD**

- Randomized controlled trial (RCT)
  - Considered the strongest design
  - Yields the best (unbiased) estimate of the counterfactual

- RCTs can receive WWC’s strongest strength of evidence rating:
  - “Meets Group Design Standards Without Reservations”
WHY ARE RCTs THE GOLD STANDARD?

• Creates a control group that is statistically identical to the treatment group except for one difference
  • Treatment group experiences the intervention
  • Control group doesn’t

• As close as we can get to comparing treatment participants to themselves!

• Can attribute meaningful outcome differences between treatment and control groups to the intervention

• Key point: By design, an RCT controls for all the observable and unobservable characteristics on which treatment and controls may differ and that might be correlated with/contribute to differences in outcomes (e.g., motivation, innate ability)
WHY ARE RCTs HARD TO IMPLEMENT?

- The intervention has already started (prior to the beginning of the evaluation)—too late to randomize
- Developer wants certain individuals in the treatment group (e.g., faculty motivated to implement)
- Ethical concerns among potential participant organizations/individuals about denying some students
- Logistics or scheduling issues don’t easily allow for randomization (e.g., high school schedules)
PLAN FOR THE SESSION

• Designing a strong evaluation
  • Randomized controlled trial (RCT) brief review

  • Quasi-experiment (QED), a viable alternative
    • Forming a valid comparison group
      • Common Pitfalls
      • Attendee discussion
    • Establishing baseline equivalence
      • Attendee discussion
      • How QEDs are reviewed by WWC

• Questions
Quasi-Experimental Design—A Viable Alternative

• When an RCT is not possible, next most rigorous evaluation that can produce evidence of the effectiveness of an intervention is a quasi-experimental design (QED)

• Aim of a QED is to mimic an RCT by implementing strategies to construct statistically equivalent treatment and control groups.

• QEDs usually depend on
  • Matching on baseline characteristics to construct treatment and comparison groups that share similar observed characteristics that are related to selection of treatment participants and outcomes (“confounders”).
  • Statistically controlling for any remaining observable differences between the two groups.
QUASI-EXPERIMENTAL DESIGN

• QEDs can only account for the baseline variables used in matching and statistical models for estimating effects

• QEDs cannot be assumed to control for all confounders
  • Confounders = characteristics related to selection into condition and outcomes
  • Some confounders may be unobservable
  • Some confounders (albeit observable) may be unavailable to the evaluator
  • Therefore, we almost never have a fully articulated model of confounders to guide our matching and estimation of effects
QUASI-EXPERIMENTAL DESIGN LIMITATIONS

• Limitation of QEDs reflected in highest potential WWC evidence rating

• QEDs cannot rule out all confounders
  • There remains the chance that they influence estimates of intervention effects
  • For this reason, the strongest evidence rating the WWC assigns to a QED is “Meets Group Design Standards With Reservations”
PLAN FOR THE SESSION

• Designing a strong evaluation
  • Randomized controlled trial (RCT) brief review

• Quasi-experiment (QED), a viable alternative
  • Forming a valid comparison group
  • Common Pitfalls
  • Attendee discussion

• Establishing baseline equivalence
  • Attendee discussion
  • How QEDs are reviewed by WWC

• Questions
 DEFINING THE RESEARCH QUESTION

- First-level decision is about the research question you want to answer—different comparison groups answer different questions
  
- Example: Is the question focused on a specific amount of exposure (e.g., one semester— at least one semester, up to three semesters?)
  
- Example: Does the sample include multiple cohorts with varying amounts of exposure? How might this affect your question --- Will different cohorts have a different number of years, semesters exposure?
MATCHING

• Matching treatment units to untreated units on key baseline characteristics is one way to construct a comparison group
  • Selection of variables for matching is generally more important than the approach to matching
  • Approaches to matching
    • Exact matching can be conducted with a small number of baseline characteristics (e.g., pretest and SES) but infeasible with more than a few variables
    • Distance matching or propensity score matching is more appropriate when matching on several variables (e.g., pretest, race/ethnicity, gender, and SES)
MATCHING TIPS

• Statistical software procedures that can implement matching methods
  
  • For a list of several software options, see following link: http://www.biostat.jhsph.edu/~estuart/propensitysoftware.html

• When matching on multiple variables, remember:
  
  • Check for balance on each matching characteristic
  
  • Control for matching characteristics in the models that estimate effects (“doubly robust estimators”)
PLAN FOR THE SESSION

• Designing a strong evaluation
  • Randomized controlled trial (RCT) brief review

• Quasi-experiment (QED), a viable alternative
  • Forming a valid comparison group
  • Common Pitfalls to Avoid when Forming Comparison Groups for QEDs
  • Attendee discussion

• Establishing baseline equivalence
  • Attendee discussion
  • How QEDs are reviewed by WWC

• Questions
AVOID USING A HISTORICAL COHORT

- Occurs when a comparison group is selected from a different time period (historical cohort)
  - For example, comparing outcomes for students participating in intervention in 2015 to outcomes of students in 2014 who did not have access to the intervention
  - Ineligible design in WWC review because events associated with a particular time period cannot be disentangled from any observed effects of the intervention
  - Does not meet WWC evidence standards
AVOID THE N=1 CONFOUND

- Occurs when only one unit is included in either the treatment or comparison group

  - For example, treatment implemented with students in college A and comparison group includes students in college B, which didn’t implement the treatment

  - Ineligible design for WWC review because cannot separate the effects of the college from any observed effects of the intervention

  - Does not meet WWC evidence standards.
AVOID COMPARISON GROUPS THAT CONSIST OF OPT-OUT

• Occurs when individuals who take up offer of the intervention are compared to individuals who were offered treatment but opted not to take it

  • Eligible design for WWC review

  • Can potentially meet WWC evidence standards with reservations

  • But concerns about unobserved differences in two groups

    • Consider characteristics related to students’ decision to opt-in as potential matching variables and covariates
PLAN FOR THE SESSION

• Designing a strong evaluation
  • Randomized controlled trial (RCT) brief review

  • Quasi-experiment (QED), a viable alternative
    • Forming a valid comparison group
    • Common Pitfalls
    • Attendee discussion

  • Establishing baseline equivalence
    • Attendee discussion
    • How QEDs are reviewed by WWC

• Questions
WWC AND BASELINE EQUIVALENCE

• For QEDs, the WWC privileges one key observable:

  • Establishing that the treatment and comparison groups are similar prior to the onset of the intervention on a pretest or other proxy measure highly correlated with the outcome

  • Less attention is paid to similarities of other observed characteristics between the treatment and comparison conditions but it is good research practice to minimize differences on them to the extent possible
**Baseline Sample Tips**

- QEDs must establish that the treatment and comparison groups are equivalent prior to the onset of the intervention (i.e., at baseline)

- Establish baseline equivalence on the **analytic sample** (sample used in analysis of impacts)
  - Baseline equivalence must be established separately for each outcome—often, the analytic samples differ

- Do not include individuals with missing pretest and/or posttest data
  - No imputation of missing data
  - Note: The loss of some individuals in the assigned sample (attrition) is not relevant in the review of QEDs
**Baseline Measures**

- Use a pre-intervention measure of the outcome (pretest of the outcome) or a proxy measure highly correlated to the outcome.

- If pretest measure not available, establish baseline equivalence with measures in two domains:
  - A continuously-scaled measure of academic achievement (e.g., high school grade point average, SAT/ACT scores)
  - Measure of socio-economic status (e.g., family income, free-reduced lunch status, Pell Grant eligibility)

- For more details, see WWC Postsecondary Education Review Protocol
ASSESSING BASELINE EQUIVALENCE

• Magnitude of the mean difference between the treatment and comparison groups

  • Effect size (Hedges’ $g$) = \( \frac{\text{Treatment mean} - \text{Comparison mean}}{\text{Pooled standard deviation}} \)

  • If ES ≤ .05, then baseline equivalence is established

  • If ES > .05 and ≤ .25, must statistically control for baseline measures in impact analysis

• Not a significance test, not a p-value
ANALYTIC MODEL FOR ESTIMATING BASELINE MEAN DIFFERENCE

• Simple t-test

• Or same analytic model used to examine impacts
  • Pretest becomes dependent variable
  • Exclude covariates
  • Include blocks
  • Include treatment indicator (to estimate mean difference)
QUESTIONS?