Randomized Controlled Trial: Design and Implementation

The E2e Project*

* We would like to thank Paula Pedro, Raina Gandhi and Karen Notsund for assistance assembling this document. For questions, contact Karen Notsund (knotsund@berkeley.edu).

1. Randomized Controlled Trial (RCT)

A. Description

A randomized controlled trial for an energy efficiency program can be designed in many ways, but the treatment and control groups must be identified at the outset. For new programs, the treatment group might be a randomly selected set of households within a utility's service territory participating in a pilot. The control group would then be households that did not participate. If the evaluation shows that a program produces net energy savings, it can be rolled out to all households.

For more targeted programs, where participants must meet qualifying criteria and funding is limited, such as a free home weatherization program, a treatment group could be determined by randomly choosing sufficient qualifying applicants to exhaust the first year of funding for the program. The control group would then consist of all qualifying applicants who do not receive the treatment. Not everyone can receive the weatherization simultaneously, so this avoids potential claims of discrimination about whose home gets weatherized first. Random assignment of implementation dates to households permits the identification of verifiable treatment and control groups. Once the trial ends, everyone can receive weatherization.

The larger the anticipated impact from a program, the smaller the sample size needs to be. The population group for the study does not need to be homogenous because the random assignment of the program will account for the differences among the participants. Statistical power tests formalize this intuition and guide decisions about minimum sample size.

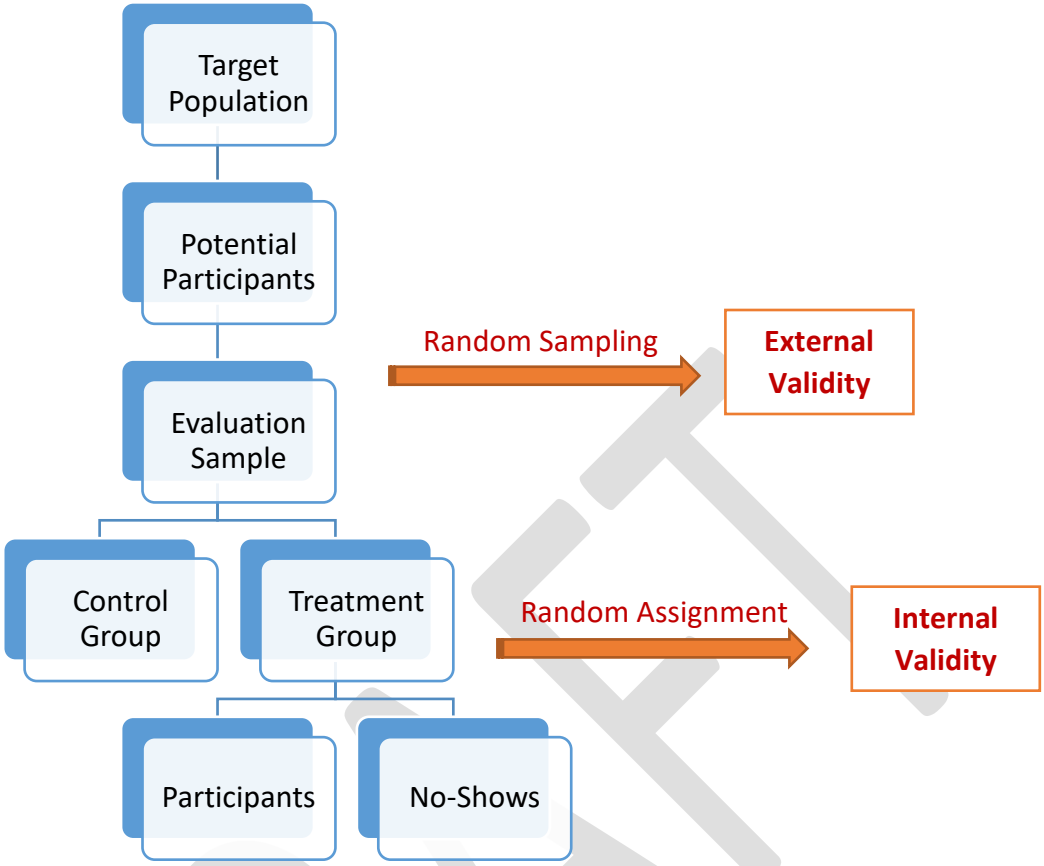

Figure 1: Basic RCT

As seen in Figure 1, potential participants are identified from a target population. If the researcher randomly selects the evaluation sample from the target population, this helps ensure that the results will be representative of the whole target population (i.e., enhances "external validity"). This sample is next randomly assigned to either the control group or the treatment group. In some cases, customers assigned to the treatment group will decide not to participate and become no-shows.

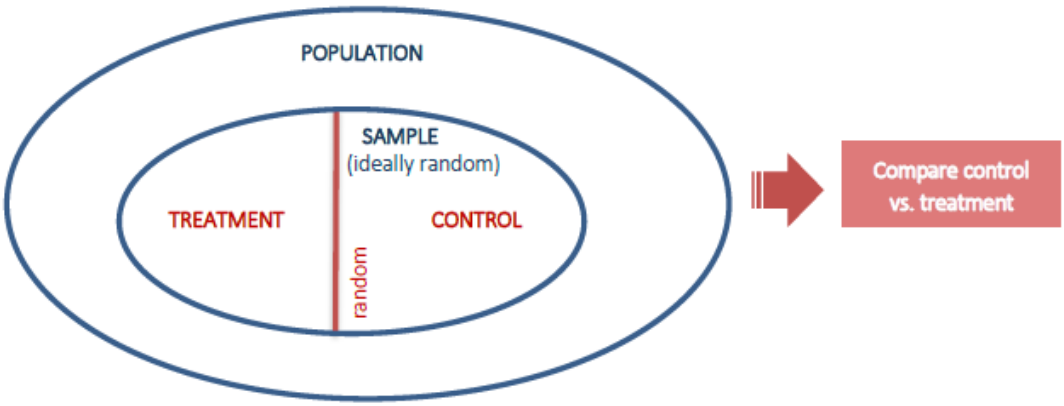

Figure 2: Creating Groups from the Population

Figure 2 diagrams the relationship between the population, study sample and treatment and control groups.

2. Design Requirements

A. Power calculations

Statistical power refers to the likelihood that an impact can be detected, when there is an impact to be detected. Power calculations determine the smallest sample size needed to detect an effect of a given size. Performing power calculations, and using the results to guide sample size, can help ensure that the proposed evaluation sample is sufficiently large to detect economically meaningful impacts. If the sample size is too small and no significant impact from the program is found, the results cannot be interpreted to suggest that the program had no impact. The program may have had an impact but with a sample size that was too small, its impact could not be detected. Performing power calculations to identify the appropriate sample size is crucial to the integrity and validity of the evaluation.

Components of power

Power is affected by several factors: variance, the effect size to be detected, and the sample size. The higher variance there is, or the "noisier" the data is, the harder it is to detect significant effects. There is little that can be done to reduce the variance aside from incorporating a baseline and attempting to control for other variables.

The smaller the expected effect of the program, the harder it is to detect a significant effect. (Note that this is true no matter what the evaluation approach.) The **minimum detectable effect (MDE)** indicates the smallest true effect that is detectable for a given level of power and statistical significance. A useful MDE is the smallest effect that would justify continuing the program.

To increase power, the researcher can increase the sample size. This makes it more likely that the evaluation finds a precise effect. Yet increasing the sample size costs money and resources, and adding more people to the sample has a decreasing effect on power. Given these diminishing returns to sample size, there is a tradeoff between precision and resources devoted to the evaluation. Thus, power calculations tell us the minimum sample needed to be likely to detect an effect of a given size (the minimum detectable effect).

Sample size and power

In considering sample size and power, it is more informative to measure unrelated people, but this is not always possible. In some circumstances, the sample can be randomized at the "cluster" level, such as schools, when students are the target population. The more correlated responses are within groups, the more this decreases the power so there must be an adequate number of groups.

If it is significantly more expensive to have people in the control or the treatment group, the proportion of people assigned to each group can be adjusted. The power calculation can incorporate the assignment ratio and determine the MDE.

B. Determining the level of randomization

In some circumstances, the evaluator may not want to (or cannot) randomize at the individual level. An alternative is **clustered randomization**, where randomization takes place across groups of individuals, such as schools or neighborhoods. Program implementation details often guide decisions on the level of randomization. For example, an evaluation of a neighborhood-level weatherization program would most likely want to randomize at the neighborhood (or higher, e.g. city) level.

The costs and benefits of clustered randomization

Benefits	Costs

• Capture spillovers
  o For example, students in a class-based intervention may share information with classmates who are in the control group.
• Compliance
  o For example, with household randomization, agents may accidentally deliver treatment information on control calls. Agent-level randomization would solve this.
• Implementation
  o May make logistics of the intervention easier.
• Reducing perceived unfairness
  o Providing the benefit of a treatment to only some households or some children in a community may be seen as unfair.
	• Decreases statistical power
	  o The more correlated measures are within a group, the greater the effect on power
	  o Requires more groups to maintain power
	• More complex statistics
	  o Must cluster standard errors
	  o Individual units or records may not be independent

C. Ensuring balance

Before beginning an intervention, it is important to check whether the treatment and comparison groups are **balanced.** After randomization, the evaluator must check that the groups do not differ on important baseline characteristics. If an evaluation with a small number of clusters, perhaps school districts for example, one group may have higher average income than the other, or be less racially diverse. Some districts may have more "green" programs, and so one group may disproportionately contain households that have been exposed to these programs. This can bias the results.

Other sources of imbalance include time-varying confounders, or if certain groups of people are more likely to drop out during the program than others. If this happens, the results must be adjusted to compensate for this.

If balance is likely to be an issue, there are alternative approaches to randomization. The sample can be stratified, dividing it into subgroups and then randomizing within each subgroup. Another technique is pairwise matching, where pairs of observations with similar baseline characteristics are identified and then assign one of each pair to the treatment group.

3. Empirical estimation

There are many statistical methods that can be used to estimate effects, depending on the type of data and corrections that need to be made. A simple, common estimation tool is called the **ordinary least squares (OLS) regression**. This tool describes a linear relationship between a response variable (such as energy usage) and an explanatory variable (number of retrofits). An OLS regression tries to fit the data with a linear model of the form: $y = \beta X + \epsilon$, where y is the response variable, X is the explanatory variable, ϵ is the error term, and β (beta) is the slope of the line.

Note that not all data are linear and should be analyzed with a linear model. Before conducting an evaluation, careful thought should be given to the type of data collected and what econometrics methods are most appropriate.

Examples of RCTs:

Wisconsin Energy Conservation Corporation

The Wisconsin Energy Conservation Corporation conducted an experiment to examine how different types of marketing and subsidies affected home audit take-up. Households were randomly assigned to different treatments, where they would receive certain marketing and subsidy materials. Separate treatment groups were exposed to versions of letters that framed the issue differently, provided different information, had different photos and themes, provided different financing information, and provided different levels of financial incentives. For example, some people were told that low-interest term financing was available, while others were offered cash incentives and the control was told that saving energy benefits everyone. Comparing audit take-up across groups provided information about what drives homeowners to conduct audits and how to best pitch audits and other programs.

Allcott, Hunt and Michael Greenstone. "Measuring the Welfare Effects of Residential Energy Efficiency Programs." E2e Working Paper #028, April 2017.

Opower

Another RCT done with Opower sought to understand how people react to social comparisons about electricity use and energy efficiency tips. The research questions were: do people eventually habituate to these reports, and after the treatment ends, how persistent are the effects? The study was done by randomizing households in the three longest-running Opower programs into a control group and two treatment groups, one receiving the reports for two years and the other receiving the reports indefinitely. For all three sites, the study showed that after treatment ends, the effect weakens but does not entirely evaporate, implying significant post-program savings.

Allcott, Hunt and Todd Rogers. "The Short-Run and Long-Run Effects of Behavioral Interventions: Experimental Evidence from Energy Conservation." **American Economic Review*, 2014, 10, 3003-3037***.

Here are several useful references on designing and implementing RCTs:

Raina Gandhi, Christopher R. Knittel, Paula Pedro and Catherine Wolfram, "Running Randomized Field Experiments for Energy Efficiency Programs: A Practitioner's Guide," Economics of Energy & Environmental Policy, 5(2), July 2016.

Gertler, Paul, Sebastian Martinez, Patrick Premand, Laura Rawlings and Christel Vermeersch, "Impact Evaluation in Practice," Second Edition, World Bank Press, 2016.

JPAL – Abdul Latif Jameel Poverty Action Lab – Introduction to Evaluations

Annika Todd, Steven Schiller, Elizabeth Stuart, and Charles Goldman, Lawrence Berkeley National Laboratory, Evaluation, Measurement, and Verification (EM&V) for Behavior-Based Energy Efficiency Programs: Issues and Recommendations, May 2012.