Research Design

> *We can randomize the allocation of an entire program...We can randomize different components of a program...We can design our evaluation to precisely test a theoretical concept that economists have worried about for years or to test a contentious policy question. In other words, we can tailor the evaluation to answer exactly the question we need to have answered.*
>
> *– Rachel Glennerster and Kudzai Takavarasha, Running Randomized Evaluations: A Practical Guide.*

Once we have established that the question can be answered through an RCT, and that it is worth the cost of answering, we then must determine the following:

1. Target population: What is the target population of this program?
2. Outcomes: What outcomes do we care about and how do we measure them?
3. Sampling: How do we construct our research sample? Should it be representative of certain groups in the population and if so, how do we ensure this?
4. Randomization Design: Who (how many) will receive each intervention and/or control, and when?
5. Power Calculations/Sample Size: For how many people do we need to measure the key outcomes?

For a broad overview of how to design an RCT, please refer to Duflo, Esther, Rachel Glennerster, and Michael Kremer. Using Randomization in Development Economics Research: A

Toolkit. In T. Schultz and John Strauss, Eds., Handbook of Development Economics. Vol. 4. Amsterdam and New York: North Holland, 2008.

For a less technical, practical guide that includes advice on the implementation as well as design of an RCT, please refer to Glennerster, Rachel, and Kudzai Takavarasha. Running Randomized Evaluations: A Practical Guide. Princeton: Princeton UP, 2013.

The Goldilocks Toolkit developed by Innovations for Poverty Action contains a number of resources for organizations that may not be able to conduct an RCT, but still want to develop strong monitoring and evaluation practices.

Target Population and Outcomes

The target and outcomes we care about should be defined by our research question. For the target population, we need to answer a number of questions: Who are the direct and indirect beneficiaries of our program? Who are the ultimate beneficiaries if we are to scale the program up? For whom would we ideally want these results to be applied to? The details of how to measure outcomes on our target population are covered in the next section on measurement and data collection.

Sampling

In some cases, the sample may be the entire target population in our research study. For example, if an evaluation targeting 3rd graders takes place in 100 schools, we may be able to obtain exam results from all students in those schools. There is no need to randomly sample which

students we wish to survey or test. However, in more cases than not, we will randomly sample respondents. In such cases, many of the same questions come up with sampling as with random assignment.

- The STEPS Sampling Size Calculator and Sampling Spreadsheet contain useful step-by-step guides to determining a sampling frame. For more information, refer to the WHO STEPS website.
- The US Center for Disease Control Provides a Guide to random sampling. The document is available on the CDC website.

Randomization Design

Randomization in concept can be quite simple; a randomized result can be generated by tossing a coin, or using a random number generator. There are, however, often many possible ways to assign treatment and control status. The main questions we need to ask are:

1. What is the level/unit of randomization? (Are individuals randomized to treatment and control conditions? Or do we randomize groups of individuals, such as entire schools or villages?)
2. What happens to the control group? (Do we deny them access permanently, or just during a certain period? Do we allow access to the control group as well, but simply try to induce greater take-up in one group ("encouragement design")?)
3. Do we want to measure spillovers? Even if we don't want to measure them, might they affect our estimates? For a comprehensive explanation of how to design experiments to measure spillover effects, see this paper by Sarah Baird, Aislinn Bohren, Craig McIntosh, and Berk Ozler.
4. Do we limit the research sample to some group that is on the border of eligibility?

5. Would we accept a design in which one individual has a higher probability of being assigned to, say, the treatment group than another individual?
6. Should we stratify, how much, and by what variables? For a discussion on the benefits and details of stratification, see this Guido Imbens' thoughts on experimental design for unit and cluster randomized trials.
7. If we don't have a complete list of units at the beginning, can we randomize as we go?
8. How should treatment status be communicated to participants: public lottery, in private, or in some other fashion?
9. If using software, which software should we use?
10. Should our randomization be reproducible so others could use our code and come up with the same allocation?
11. If we're not satisfied with the level of balance between our groups, can we re-randomize? For a comparison of the relative merits of ex-ante stratification, pairwise matching, ex-post re-randomization to achieve balance, etc., see David McKenzie and Miriam Bruhn's paper on various randomization strategies. A Development Impact blog post by the same authors goes into the mechanics of stratification for balance further.

Power Calculations/Sample Size

An experiment must be sensitive enough to detect outcome differences between the treatment and the comparison groups. The sensitivity of a design is measured by statistical power, which, among other factors, depends on the sample size – that is, the number of units randomly assigned and the number of units surveyed. The statistical power of a study determines the likelihood that the study will detect an impact of the treatment when there genuinely is an impact to be detected. In other words, maximizing statistical power by picking an appropriate sample size minimizes the likelihood

of committing a Type II error: rejecting the proposition that the treatment has an impact when it does indeed have an impact.

- Owen Ozier's slides on sample size and power calculations, originally developed for an IPA-J-PAL training in 2010, are a useful introduction to the determinants of power and the relationship between sample size and power.
- The World Bank's Development Impact blog has had a number of interesting posts on power calculations; refer to the "power calculations" tag on the blog for more details.
- For simple sample-size calculations in designing cluster randomized trials, see this paper by RJ Hayes and S Bennett in the International Journal of Epidemiology.
- Optimal Design is a free software program for conducting power calculations, compatible only with Windows operating systems (for advice on running Optimal Design on Macs, see here.) The program and documentation are available for download here. The document available here contains exercises and a step-by-step guide that demonstrate how varying parameters can affect the statistical power of a study.
- The IFS has a useful practical guide to power calculations, including sample Stata code for power calculation simulations.

Please note that the practical research resources referenced here were curated for specific research and training needs and are made available for informational purposes only. Please email us for more information.

---

https://www.povertyactionlab.org/research-resources/research-design