GOLDILOCKS TOOLKIT

Impact Measurement with the CART Principles

FEBRUARY 2016

Acknowledgments: The Goldilocks toolkit was authored by Mary Kay Gugerty, Dean Karlan, and Delia Welsh, with editing and design by the IPA Communications Team (David Batcheck, Laura Burke, Jennifer Cowman, Heidi McAnnally-Linz, Megan McGuire).

ipa
INNOVATIONS FOR
POVERTY ACTION

Impact Measurement with the CART Principles

For many organizations, a central goal of monitoring and evaluation is to prove that programs are making a difference—that they have an *impact*. Not only is it important to know if programs work, but providing hard proof can attract much-needed funding and may also improve an organization's reputation. The reality, though, is that not everyone can and should measure impact: sometimes it's not possible to muster up a sample size that would be large enough to conduct a good study, or there simply isn't anything to randomize. When programs are structured such that impact measurement *is* possible, it's still important to approach evaluation carefully. Impact measurement that aligns with the CART Principles must be well-designed, implemented well, and timed appropriately. If the evaluation design or fieldwork are sub-par, results will be biased (meaning *wrong*), which can lead organizations and policymakers to start or continue programs that have little or no impact, or to miss opportunities to expand effective programs.

In this section, we'll go through the common biases that get in the way of measuring impact and explain what it means to conduct credible, actionable, responsible, and transportable impact evaluation.

The Problem: A Proliferation of Bad Evaluations

Over the last decade or so, "impact" has become a buzzword in the nonprofit sector, and organizations are increasingly encouraged to measure and demonstrate impact. The trend to measure impact has brought about an increase in randomized evaluations but also many inappropriate methods for measuring impact. While probably well-intentioned, these evaluations can produce biased, misleading results and represent a waste of resources. Why does bias matter? Bias clouds our ability to see the true impact of the program, leaving us back where we started: with little usable information on impact. But it can even leave us worse off, with fewer resources and little to say for it.

In evaluations, there are three main sources of bias to be aware of:

Bias from external factors

Some factors in the world, such as weather, macroeconomic conditions, or other government or NGO programs, are highly likely to influence the outcome of interest. If both the group that receives the program and the comparison group are not both exposed to these same outside factors, bias is likely to occur.

Bias from self-selection

Unobservable characteristics about those who choose to participate, versus those who do not, may influence who chooses to participate in a program. If you compare those who choose or "self-select" to participate and compare them to those who did not choose to participate, you may end up with biased results because your two groups are composed of people who are fundamentally different. It is important to recognize that many characteristics, like spunk, desire, and entrepreneurial drive are quite hard if not impossible to measure (except through observing whether they participate!).

Bias from program selection

Program selection bias is similar to self-selection bias, except the problem is driven by the program allowing certain people to participate rather than others. As an example, suppose an NGO started an agricultural input provision program and began the program in the most agriculturally productive area of a country. Comparing harvests of that program to harvests in another area will yield biased estimates of impact by overestimating the potential for yield increases in the rest of the country.

To overcome these biases and determine that there is a causal link between the program and the changes in outcomes that occurred, organizations need a way to measure what would have happened if a program had not existed. They need a reliable estimate of the *counterfactual*. That brings us to the first CART principle, Credible.

Credible

To credibly measure the impact of a program, we need to compare the outcomes of a program to what would have happened otherwise. It may seem like wishful thinking to try to know *what would have happened otherwise*. In our daily lives we are never able to know would have happened if we had done things differently. But in research, we are able to design evaluations that allow us to compare the actual to the "what if" with a high degree of confidence. Conducting impact evaluation with a valid estimation of the "otherwise" (called the counterfactual) enables us avoid the biases described above. These methods include randomized controlled trials and quasi-experimental methods.

Randomized Controlled Trials

Commonly considered the "gold standard" in research,[1] randomized controlled trials, or RCTs, refer to a study design that involves randomly assigning a large number of individuals (or households, communities, or other units) to either a treatment group that receives an intervention or to a control group, which does not. As long as the assignment is truly random, we will have two groups that are similar on average both in observable characteristics (e.g. gender, income, ethnicity) and unobservable characteristics (e.g. self-motivation, moral values).[2]

With two groups that are similar on average, we can measure the impact of a program by measuring the difference over time between the group that received the program and the one that did not.[3] Over the course of the study, the only difference between the two groups should be the effects of the program itself. To learn more about designing an RCT, we recommend reading *Running Randomized Evaluations* by Kudzai Takavarasha and Rachel Glennerster.

[1] The Institute of Education Sciences (IES) and National Science Foundation (NSF), National Academy of Sciences, Congressional Budget Office, U.S. Preventive Services Task Force, Food and Drug Administration, and other respected scientific bodies consider RCTs the strongest method of evaluating the effectiveness of programs, practices, and treatments.

[2] Which one can verify using observable information available prior to the study or data available at the end of the study, but that does not change or is easily recalled without bias, such as gender and marital status.

[3] Notice the key difference between random assignment and random sampling, in which the 'treatment' and 'control' groups are selected after the program begins. Many organizations conduct an evaluation by randomly sampling in an area that does not receive the program as the comparison. This is not the same as a randomized study.

Quasi-experimental Methods

Apart from RCTs, other methods can credibly measure the impact of programs under the right conditions. With these “quasi-experimental” methods, rather than setting up the treatment groups beforehand through random assignment, researchers use different techniques to build treatment and control groups.

Although quasi-experimental designs can be just as valid as RCTs, these methods require more assumptions, institutional facts to justify the analysis, and often require researchers with expert statistical and econometric analysis skills to crunch the numbers. Often, these evaluations also require a large amount of data collection. For all these reasons, the following methods should be used to measure impact with caution, expert guidance, and only when an RCT is not possible.

- ***Matching:*** In matching, researchers attempt to find a comparison group that closely resembles those who participated in the program based on various observable characteristics. Common matching characteristics include gender, income, education, and age. The overall validity of this approach depends on how well these *observable* characteristics allow researchers to create two similar groups of people, as well as how important unobservable factors are to program outcomes – which can be hard to know in advance.
- ***Regression discontinuity:*** This method involves using an existing eligibility cutoff for a program to help create treatment and control groups that are just above and just below the cutoff line. The eligibility cutoffs for many programs, such as college admission or social welfare programs, are set in absolute terms and individuals just above the cutoff and just below are often very similar. The challenge is that it can be hard to know at what point beyond that cutoff those similarities end.
- ***Difference-in-differences:*** This method measures changes in outcomes over time (before and after) of the program participants relative to the changes in outcomes of non-participants. Hence it combines a before/after comparison with participant/ non-participant comparison. A critical assumption is that in the absence of the program the difference between participants and non-participants would stay the same over time – often a big assumption.

Actionable

Creating an actionable impact evaluation means designing a study that will generate useful evidence, and making sure the organization intends to use the evidence, regardless of the results. “Actionability” refers to an organization’s willingness to change based on the results. If an organization will not change a program regardless of the results of an impact evaluation, they should save their money and not conduct the evaluation in the first place.

Impact evidence to guide actionable decisions can be generated in a number of ways. To guide program design, an organization might test different elements of a program, rather than only evaluating the whole program, which typically includes multiple interventions. Evaluating different program elements allows managers to identify which elements are most effective, even though they may not measure overall impact.

Additionally, timing an evaluation with the project cycle is key to translating results into action. If you evaluate a program during a pilot stage (but after a proof of concept), for example, it will allow you to fine tune the program for optimal effectiveness before rolling it out on a large scale. It can be difficult to change a program after roll-out, when funders are invested. Managers at that stage are often hesitant to change gears or consign the program to the scrap heap.

Responsible

Given the high costs of collecting and analyzing data, organizations have to weigh the costs and benefits of conducting an impact evaluation against those of *not* doing the evaluation. These costs and benefits have monetary as well as societal components.

Although randomized evaluations can be expensive, that cost can be well worth it when the evidence is used to help a program and others around the world improve effectiveness. If the knowledge gap is large, the benefits have the potential to be large as well. We should also note there are ways to conduct RCTs at a low-cost, such as by introducing random assignment to new initiatives as part of program roll-out or by using existing administrative data to measure key outcomes instead of collecting original data. These designs can save organizations lots of money, making them a responsible use of resources.[4]

The social costs refer to the ethics of an evaluation. Some have argued that randomized evaluations are unfair because they withhold a “treatment” from a portion of study participants. However it is an extremely rare situation when a program or organization has the resources or ability to roll out a program to *everyone*. In most cases, some people are excluded, so in that sense, an RCT is no different. Second, at the beginning of the evaluation we do not know if the intervention is effective, so we cannot argue that an RCT helps some and not others. Rather, there is a social cost of not knowing whether a program works or not, and continuing to waste precious resources if they don’t.

Transportable

Some argue that every program operates in a unique context, and that just because a program works well in one context does not mean that it will work in another. That’s true to some extent–just because a job training program worked well in Bangladesh doesn’t mean the same program will work equally well in Peru. The ability to apply evidence to other contexts requires more than an estimate of impact. It requires understanding *why* the program worked and the conditions that need to be in place to get the same result. Transportability, or the extent to which evaluation results can be used in other settings and contexts, begins with the evaluation design.

Impact evaluations should be designed from the start using a theory of change that tests a concrete hypothesis. If the results of the impact evaluation support that theory, then the initial findings may also to be validated through further study to be sure that the mechanism works before implementing the program more widely.

[4] Before only collecting data at endline, you have to be sure that your randomization was effective and that the two groups are the same on average. If you have any questions about whether the randomization was indeed effective, collect information at baseline.

The degree of transportability is what helps determine which programs are scalable and in which contexts. The health sector has many examples of programs that might work in multiple contexts. Since many treatments (such as vaccines or immunizations) are effective regardless of context, once the treatment has been found, it can be scaled to other areas quickly (assuming there are no adverse side effects).

The goal of transportability is to make sure that all impact evaluations generate evidence that will help others design or invest in better and more effective programs, not just get a stamp of approval for their own program. That way policymakers will have better evidence to implement programs in the future.

Decision Time

When deciding if an impact evaluation makes sense, donors and nonprofits should begin by thinking about whether or not an impact evaluation would meet the CART criteria. A good starting point is to ask the following key questions:

1. Are we able to credibly estimate the counterfactual? (Is it credible?)
2. Are we committed to using with the results, whether they are positive or negative? Is the timing right? (Is it actionable?)
3. Do the benefits outweigh the costs? Would this evaluation add something new to the available evidence? (Is it responsible?)
4. Would the evaluation produce useful evidence for others? (Is it transportable?)

If you can answer “yes” to these questions, you may be ready for an impact evaluation. If the answer to any of the above questions was “no,” impact evaluation may not be appropriate for your program at this time. In that case, you should focus on developing systems to collect monitoring data that deliver information to promote learning about how to best implement programming.

Sources

Banerjee, A., Cole, S., Duflo, E & Linden, L. (2007). Remedying Education: Evidence From Two Randomized Experiments in India. *Quarterly Journal of Economics*. 122(3): 1235-1264.

Coalition for Evidence-Based Policy. (2013). Demonstrating How Low-Cost Randomized Controlled Trials Can Drive Effective Social Spending: Project Overview and Request for Proposals. Available at: http://coalition4evidence.org/wp-content/uploads/2014/02/Low-cost-RCT-competition-December-2013.pdf

Department of Health and Human Services. Vaccines are Effective. Available at: http://www.vaccines.gov/basics/effectiveness/

Bhattacharya, D., Dupas, P., & Kanaya, S. (2013). Estimating the Impact of Means-tested Subsidies under Treatment Externalities with Application to Anti-Malarial Bednets**.** NBER Working Paper. Available at: http://www.nber.org/papers/w18833

Glennerster, R., & Takavarasha, K. (2013). Running Randomized Evaluations. Princeton University Press.

Karlan, D.,Morten, M., & Zinman, J. (2012). A Personal Touch: Text Messaging for Loan Repayment. NBER Working Paper. Available at: http://www.nber.org/papers/w17952.pdf

Roodman, D. (2011). Bimodality in the Wild: Latest on Pitt Khandker. Daniel Roodman's Microfiance Open Book Blog. Available at: http://www.cgdev.org/blog/bimodality-wild-latest-pitt-khandker

Glewwe, P., Kremer, M., & Moulin, S. (2002). Textbooks and test scores: Evidence from a prospective evaluation in Kenya. BREAD Working Paper, Cambridge, MA.