ECONOMIC GROWTH CENTER
YALE UNIVERSITY
P.O. Box 208629
New Haven, CT 06520-8269
http://www.econ.yale.edu/~egcenter/

CENTER DISCUSSION PAPER NO. 973

Cairo Evaluation Clinic: Thoughts on Randomized Trials For Evaluation of Development

Dean Karlan
Yale University
Innovations for Poverty Action
Jameel Poverty Action Lab

June 2009

Cairo Evaluation Clinic:
Thoughts on Randomized Trials for Evaluation of Development

Dean Karlan

Abstract

We were asked to discuss specific methodological approaches to evaluating three hypothetical interventions. This article uses this forum to discuss three misperceptions about randomized trials. First, nobody argues that randomized trials are appropriate in all settings, and for all questions. Everyone agrees that asking the right question is the highest priority. Second, the decision about what to measure and how to measure it, i.e., through qualitative or participatory methods versus quantitative survey or administrative data methods, is independent of the decision about whether to conduct a randomized trial. Third, randomized trials can be used to evaluate complex and dynamic processes, not just simple and static interventions. Evaluators should aim to answer the most important questions for *future* decisions, and to do so as reliably as possible. Reliability is improved with randomized trials, when feasible, and with attention to underlying theory and tests of *why* interventions work or fail so that lessons can be transferred as best as possible to other settings.

Keywords: program evaluation, randomized control trial

JEL Codes: B41, O12, H43, J08, H54, D73, D12

1. Introduction

Why do we evaluate? Three reasons stand out: to know where to spend limited resources, to know how to improve programs, and to motivate those with money to give or invest more.

I would like to begin with a thought experiment from the utilitarian philosopher Peter Singer. Would you save a child drowning in a lake if it would cost you $100 in ruined clothing or a missed appointment? Most people answer yes to this question. But would you also send $100 right now to an NGO in a poor country to save a child? Many people say no, arguing that no one really knows if their $100 can save a child or will just get wasted. This is a common excuse for *inaction*. Evaluation rebuts this excuse.

There has been much discussion about the use of randomized control trials (RCTs) versus other methods of evaluating programs. But, in many cases, this hard split between experimental and other approaches is manufactured and masks the overlap between experimental and qualitative methods that can characterize good evaluation. In this note, I begin by outlining some common misunderstandings of the measurement method, attribution and feasibility of randomized control trials (RCTs). Then, I will describe three examples of common development programs – conditional cash transfers, infrastructure and anti-corruption measures -- and the circumstances in which RCTs should or should not be employed as part of the evaluation strategy.

2. Misperceptions of RCTs

A common misperception is that one must choose either to do a qualitative evaluation or an RCT. Underlying this is an erroneous spectrum of "attribution" rigor, with RCTs on one end and qualitative methods on the other. In reality, qualitative methodologies are not the opposite of RCTs. For one, a good RCT evaluation often involves a thorough assessment of how the program functions, its initial design, theory of change, beneficiary participation, etc.

To clarify the discussions on evaluation methods, it is imperative to separate our conversations about collecting data and measuring outcomes—what to measure, how to measure, and who to include in the process—from how to establish causality between the outcomes and intervention. RCTs establish causality by providing a measure of the counterfactual: *what would have happened* had the program or policy not existed. Just as is standard practice in medical trials, they achieve this by randomly assigning people to treatment and control groups, so that, except for the random program or offer, the groups are alike on observable and unobservable characteristics if the sample size is sufficiently large.

Establishing Causality

The random assignment is helpful because of selection bias, or in other words because program participants are often different from non-participants. If instead we were to compare those who *could* participate in a program but *choose* not to, we would end up comparing two potentially very different sets of people. It is easy to see how these groups might differ in important but hard to measure ways. Those who join the program might be more driven to improve their situation, or more empowered, or better educated. They might have more free time. Researchers often try to control for these differences, but inevitably there are omitted variables, or others, like motivation, that can be problematic to measure. These differences mean that estimates of the impact of the intervention are biased, since differences in outcomes in the treatment and control groups may result from these unobserved characteristics, rather than being caused by the intervention.

Data and Measurement

Quantitative outcome measures are useful for evaluations because they allow researchers to establish statistical significance for program impact. But RCTs do not specify any one method for data collection. Both quantitative and qualitative data can be used within the RCT framework, often in combination within the same evaluation. Methods from economics, sociology and psychology or other disciplines can be used, as well as participatory processes involving local voices (e.g., see Chattopadhyay and Duflo 2001 which found that women in West Bengal were more likely to participate in the policy making process if the leader of their village council was a woman), among others, and even "outliers" as Chambers discusses in this forum (Karlan and Zinman 2009).

A common misperception directed at advocates of RCTs is that we suggest they can and should be conducted on every program. RCTs are an important research tool because the causality they establish provides rigorous measure of program impact, and thus helps to know whether to replicate elsewhere, as well as how to improve. However, RCTs are not always feasible. Where RCTs are appropriate depends partly on the situation, and also on the question being asked. And as Ravallion's (2009) article in this forum discusses, one should never start first with the methodology and then figure out what to ask. The evaluators must first establish the questions that need answering, and then examine the most appropriate tool to answer them. When feasible, RCTs provide the most unbiased estimate of program impact, but merely being feasible by no means suggests they should be done just for the sake of doing one. Where no convenient identification strategies exist RCTs are without doubt the most practical means of creating a credible research setup.

Creative Approaches in RCTs

While we emphasize that RCTs cannot work everywhere, many settings which seem infeasible are in fact feasible with a little creativity. For example, interventions can often take advantage of implementation limits and randomize at the community or other

geographical level rather than randomly selecting individuals into control and treatment groups. There are several evaluations measuring the impact of microfinance that use this approach. In other cases, differences in the intensity of marketing a program to different areas (encouragement design) can be exploited in an RCT. The key criterion for RCTs is sample size, in separable enough units such that spillovers and general equilibrium effects can be measured. If planned properly and if the effects are not overly aggregated, (e.g., at a country level), then careful RCT designs can measure both the direct impacts of the intervention as well as the positive and negative spillovers onto groups outside of the direct beneficiaries. These are in fact some of the most exciting RCTs to read about, because they help us understand not just whether an idea works on a particular individual, but how it will play out on a larger scale with direct and indirect effects.

Static vs. Dynamic Implementation Approaches

Another common misperception of RCTs is that the intervention must be homogenous and static. Indeed, "emergent, complex" or "complex" interventions, such as those discussed by Rogers (2009) in this forum, are not more difficult for an RCT to handle than for a non-RCT. Arguments that suggest complexity and a dynamic process wreak havoc with an RCT are failing to recognize what exactly an RCT gets us. An RCT simply helps to generate an objective comparison group against which to compare changes. The intervention itself of course can be static and simple, or complex and changing. If the latter, then the evaluation is of course described as such: one is evaluating a process, an opportunity coupled with some resources, a dynamic and fluid intervention that was led in a certain way, etc. The key here is that it is the process, not the individual activities that make up the program implementation, is thus being evaluated. If the project were to work, then what needs to be replicated is the process of putting resources into place, facilitating the use of them, etc. This is much akin to many community development interventions in which resources such as training and customized technical assistance are provided to communities and facilitation exercises are put in place to help communities grow and prosper.

We are conducting just such an evaluation using an RCT approach, complete with qualitative and quantitative tools, of The Hunger Project in Ghana, and of a community-driven development program in Sierra Leone. It is important of course to understand that what is being evaluated here is a collaborative process rather than a clearly defined intervention. It is not possible to know up front what inputs the particular actors will select, nor to expect that the same process elsewhere would yield the same choices. Thus the lessons from such an evaluation are about the changes one can expect from just such a process—not from the specific choices and investments the actors choose to make, but from the process of facilitating and/or financing the villages as they develop the program themselves. That said, if program officials or managers were interested in measuring the individual impacts of the activities that make up the intervention, an RCT could be designed to deliver discrete results from complex interventions. This would require randomly varying the components of the intervention into multiple treatment groups. The likely comparison would be impact of a base set of services, with or without the interaction of one or more add-on components.

RCTs have an important advantage over other methods here, because they can address selection biases inherent in many social programs, and in addressing the direct impacts of different activities in a multiple treatment design. For instance, if one conducted an evaluation of business training and found an increase in profits, especially among those who were found to engage in better recordkeeping, does this suggest training in recordkeeping should be promoted? Maybe recordkeeping is a key component of the training, or potentially the better entrepreneurs naturally engaged in recordkeeping. RCTs can disentangle these issues by assigning participants to receive training with or without a special recordkeeping module.

Another common misperception of RCTs is that they measure impacts of an intervention only on the population average, ignoring the differential impacts on different segments of the population. In fact, given sufficient sample size and a sampling plan that includes a variety of people that might be eligible for the broader program, an RCT can help identify groups for which the program has the largest impact and groups for which the impact is insignificant or even negative. For example, one surprising result from an RCT measuring the impact of a microenterprise business training program in Peru, was that businesses who expressed no interest in additional training actually benefited somewhat more from the program (Karlan and Valdivia 2008).

3. Three examples

Ravallion's article in this forum provides an excellent overview of the types of questions one must ask in the beginning of the evaluation process in order to define the aim and scope of the evaluation, and thus the key research questions. As he discusses, depending on the unit of assignment, randomization will or will not be feasible. These three examples provide an excellent spectrum of just that point. I will discuss here both broad plans for each on how one could evaluate them, and then specific ideas within each on how subsidiary questions about specific implementation questions could be answered through randomized trials, even if the core intervention is employing other methods to assess its overall impact. These ideas are not in lieu of the overall non-experimental evaluation, but can provide useful methods of generating precise and objective data to help with important future implementation questions.

Turning to the first example, conditional cash transfers (CCTs), the recommended method for impact evaluation is the randomized control trial, involving quantitative and qualitative data collection. These have been conducted in several countries. Where governments have had limited resources to scale up CCT programs randomization is an especially fair and transparent way to distribute benefits in a staged manner. Recent research has shown that, designed appropriately, CCTs can be an effective means to achieve important public policy goals. However, the question of how best to implement these programs is far from settled. For example, implementation questions include how

frequent to make the payments, whether to consider adding savings services, and whether to coincide payment with education expenditures.[1]

For the second example, infrastructure, there are several options for designs that are technically feasible, but that require a varying degree of commitment on the part of government officials managing the programs. I will discuss port rehabilitation, trunk roads and rural feeder roads. Unfortunately, evaluators are too often asked to evaluate only after it is too late. Regardless of the method employed, it is far preferable to set up the evaluation in advance, have clear objectives and be inclusive about what and how to measure the results.

The evaluation of port rehabilitation and trunk roads could involve a heavy focus on process evaluation methodologies. The first step will be to establish a log frame, with targets for example for the number of days wait time and the number of days to transport; the cost of shipping and transporting over land; the value of goods being shipped; the quantity of goods shipped; and the number of ships, trucks and cars entering or leaving. There is potential for econometric tools to be used, depending on differential effects on industries, and tariffs, for example. This is simply program monitoring, and it is important both for implementation management and accountability for results.

In these cases, RCTs can be employed to help answer critical aspects of the theory of change for the program, but are not likely to involve the entire intervention. For example, a key question for a port rehabilitation in a developing country might be "Will lower transport costs lead to more growth of industry in rural areas?" In this case, one could consider an RCT which randomly subsidizes transport costs in some areas, to examine the change in economic activity as a result.

Answering key policy questions on the impact of feeder roads programs via an RCT approach can be technically feasible, but is also likely to require a great deal of commitment on the part of policy makers. This example is one that has huge benefits in terms of policy lessons for other countries, but is also one that we recognize might be difficult to accomplish politically. If there are enough roads, and geography and construction costs permit, there is potential for a randomized phase-in of the road construction. Imagine a ten-year plan to improve or build rural feeder roads. Randomizing the order is both (a) fair, and (b) easily evaluable. This could be implemented incorporating road prioritization within the ten-year plan, if some roads are more important for economic and geographic (or political) reasons. Enterprising policymakers might recognize that one advantage of an RCT in this context is that it avoids political favoritism to decide ordering. (That is, roads would be selected for wave 1 or wave 2 in a transparent deliberative process, then the order of road construction within each wave would be randomly drawn, hence fair.) In this case, the RCT is one facet of the evaluation design. It could also involve the use of econometric methods, including difference-in-difference approaches, before versus after, or a cross-sectional

[1] For a good discussion of how to design choice environments that help people choose ethically, see *Nudge* by Richard Thaler and Cass Sunstein.

comparison of built versus not built (for example towns 5 miles from the repaired or built road, versus 5 miles from unrepaired or unbuilt road).

The final example, anti-corruption measures, is not usually the place to look for attribution-style evaluation, although there can certainly be some measurable process outcomes such as arrests made, or politicians thrown out of office. But inside the box of how and why public officials resort to unlawful tactics, much can be learned. And furthermore, this is one area where transparency in the evaluation approach really matters! For example, Brazil's municipal audits were televised. Work by Ben Olken (2007) in Indonesia is another good example, enabling us to learn the relative effectiveness of competing anti-corruption methods, through a mixture of qualitative (perception of corruption and participatory methods from village meetings) and quantitative (quality of actual roads) data collection.

4. Conclusion

One final advantage of the RCT approach is the independence that it allows, in that one can establish clear statistical tests ex-ante, and then let the data speak as to whether something worked or not. Ultimately, the goal for evaluation should be to help decide what to do in the future. This is both for donors who need to know where to put their money, for skeptics who want to see that programs can work, and for implementers who need to know how best to design their programs. Some of the most exciting work uses mixed methods by incorporating qualitative methods into randomized trials, and by using randomized methods for evaluating dynamic and complex processes, such as community development programs.

In this paper, I have focused on a couple key issues in the debate surrounding impact evaluation methods: the parsing of *what to measure* versus *what to compare*. Looking at these distinct questions we can see RCTs focus on the latter, and are flexible to including many participatory, qualitative, and quantitative methods for the former. I have also tried to dispel some common misperceptions about the extremes of the debate. Even proponents of RCTs do not advocate that they be conducted everywhere and for every program. If I were to hazard a guess, it would be that less than 1 percent of evaluation budgets are used for RCTs. I think they should be done more, but not 100 percent (or 99 percent) either.

References

Chattopadhyay, Raghabendra, and Esther Duflo. 2001. *Women's leadership and policy decisions: evidence from a nationwide randomized experiment in India*. MIT.

Karlan, Dean, and Martin Valdivia. 2008. Teaching Entrepreneurship: Impact of Business Training on Microfinance Institutions and Clients. *Yale University Economic Growth Center working paper*.

Karlan, Dean, and Jonathan Zinman. 2009. Expanding Credit Access: Using Randomized Supply Decisions to Estimate the Impacts. *Review of Financial Studies.*

Olken, Benjamin. 2007. Monitoring Corruption: Evidence from a Field Experiment in Indonesia. *Journal of Political Economy* 115: 200-249.