Randomized control trials are the best way to measure impact of microfinance programs and improve microfinance product designs.

Authors: Karlan, Dean; Goldberg, Nathanael; Copestake, James

Source: Enterprise Development and Microfinance, Volume 20, Number 3, September 2009 , pp. 167-176(10)
Publisher: Practical Action Publishing

Dear James,

As you and readers of this column are aware, the microfinance industry is rife with debates about the best way to reach and serve the poor. Many issues have been eloquently argued in these pages, on listserves, and at conferences. The exchanges, though, rarely provide clear resolution because contributors lack the hard evidence needed to settle arguments. Some of the questions are philosophical, but many are plainly empirical, and thus answerable in a more objective fashion. Practitioners, donors, and policymakers need answers, not more debates. This is where research can provide tremendous value. Consider some previous Crossfire debates:

- 'Informality is a constraint on economic growth; greater efforts should be made to help microenterprises to register'
- 'The provision of support to HIV/AIDS-affected clients jeopardizes the sustainability of MFIs'
- 'Most BDS programmes are wasting their time attempting to make survivalist enterprises grow'

We suggest that the best way to answer these questions, and many more, is through randomized control trials (RCTs). We arrive at this conclusion after years of following the central debates in microfinance, including: Is it better to serve fewer, poorer, clients, or large numbers of the "near poor"; should we focus on credit, savings, and insurance only, or the expanded services called "microfinance plus"? These questions are still open and will remain so as long as industry leaders continue to make arguments and policy based on weak evidence—or none at all.

The recent BDS Crossfire is a telling example: while it was being argued, we were using RCTs in the field to collect evidence about the effectiveness of BDS programs. With FINCA Peru we randomly assigned village banks to receive business training, or to remain as a standard borrowing group (the control). We found that the training improved profits for both clients and the MFI.

RCTs—where they are feasible and practical—are the best method for measuring impact because they allow us to estimate what would have happened without the intervention under study. Measuring the counterfactual—"what would have happened"—is essential for establishing causality. Without it, we have no way of knowing whether changes in participants' lives are caused by the program, by outside factors, or, most problematic, by unmeasured characteristics of the participants themselves. As in medical trials, we isolate the impact of an intervention by randomly assigning subjects to treatment and control groups. This makes it so that all those other factors which could influence the outcome are present in both treatment and control, and thus any difference in outcome can be confidently attributed to the intervention.

The random assignment to treatment and control is best because it minimizes the number of fancy and questionable assumptions needed to generate an unbiased impact estimate. Without random assignment, if for example we were to simply compare borrowers to non-borrowers, we have to assume that those who borrow are not more driven, more entrepreneurial, more ambitious, etc., than those who choose not to borrow, or that lenders choose to enter into villages serendipitously, rather than target communities that lack good credit services or that have good, growing markets. These are huge, typically unrealistic, assumptions.

RCTs also offer the ability to test product designs, not just impact of credit versus no credit. We are using them to test the impact of savings-led microfinance, of optimal loan pricing, of SMS reminders for savers, and of credit itself. In each case the experimental process is similar: we offer a new approach to a random subset of clients (or villages), leaving others to serve as a control group, and we compare their outcomes. We can measure impact on household income, women's empowerment, mental health, child labor, etc., and MFI profits too.

We close by emphasizing that RCTs cannot be done in all circumstances. There are lots of Crossfire examples where RCTs would be an inappropriate evaluation strategy. Take a recent example, 'Will electronic banking reach the very poor?' To answer that one should compare users' poverty levels to the community as a whole. No RCT required. But RCTs could help find ways to *improve* reaching the poor, e.g., by randomizing pricing or marketing strategies.

The bottom line is that most programs are *not* evaluated using randomized techniques, and *more* (but not *all*) should be. Well-implemented RCTs provide particularly powerful ways to measure impact or improve product design because they guarantee the best unbiased estimates of program or product design impact. It is time to stop speculating and start collecting rigorous evidence about what works and why, so that we can maximize the impact of microfinance programs around the world.

Yours,

Dean and Nathanael

Dear Dean and Nathanael

I agree there is scope for better research into microfinance impact and how it can be improved. But I am worried that excessive enthusiasm for RCTs could lead to the neglect of other problems and solutions. Hence I'm looking forward to learning of examples that demonstrate its relative effectiveness and to discussing when and how it can be used alongside or instead of other methods. These lessons won't have to be based on randomized control trials, but I do hope you will be able to offer some cost-benefit estimates, no matter how rough.

Let me start by agreeing that when it comes to internal validity, RCTs can address the attribution problem in a neat way and generate clearer findings. I think they have other potential strengths also. Because they require the client organisation to make an operational decision they force researchers to engage more closely with its senior management, and the potential inequity of randomly assigning a treatment to only some clients should stimulate more serious discussion of research ethics. They can be used to investigate the effect of particular ‘treatments’ on a wide range of different outcome indicators simultaneously, illuminating trade-offs. Finally, given weak incentives to do so, I also welcome almost anything that encourages more investment in empirical evaluation of donor financed interventions and strengthens evidence-based policy-making.

However, I also have four concerns. The first of these is ***problem selection bias***. I am worried by reports of bright young “*randomistas”* narrowing the research agenda by selecting issues for research to fit their preferred tool, rather than finding the best tool to fit the most important issues. For example, use of RCTs to test product design changes should not divert attention from other influences on impact that may be harder to randomise: geographical targeting methods and organisational culture, for example. We should not let academic preferences for novelty divert resources away from research that is less cutting-edge but operationally more important - finding out why many clients leave, for example. More generally, perhaps the most important research on microfinance impact is sector-wide and naturalistic in scope: addressing how the financial system operates without regard to the interests, actions and options of particular agencies.

My second concern is with ***external validity*****.** RCTs require a fixed investment and generate evidence at the end of a discrete period of time, rather than continuously. This accentuates the difficulty of choosing which few among many possible ‘treatments’ should be studied, where and when. The value of findings then depends upon their transferability. I admire your work with FINCA Peru. But how relevant are your findings on bundling BDS and microfinance to the very different contexts of Bangladesh, Botswana or indeed Piura? Even scaling up use of a ‘proven’ innovation within the same organisation entails significant contextual change. You seem willing to make very strong assumptions about the universality of product performance across contexts.

Third, there is the issue of ***cost effectiveness***. I’m very much in favour of experimentation and testing, but remain to be convinced that RCTs are necessarily the most cost-effective way for managers and policy makers operating in complex, diverse and uncertain contexts to evaluate them, compared to triangulating routine monitoring data against focus group discussions and individual satisfaction surveys, for example.

Fourth, there may well be other more technical problems with RC studies. For example, it will not always be possible to ensure that treatment and control groups are not contaminated through ***spillover effects*** between them: response to not having a treatment being affected by my knowledge that others are having it, for example.

More generally, some enthusiasts for RCTs seem to reveal a limited grasp of the philosophy of science by implying that their approach (like medicine) is somehow ‘hard science’ whereas other methods are ‘soft.’ I look forward to more precise illumination of how, when and where the approach can usefully add to microfinance research, institutionalised learning and innovation.

Yours

James

Dear James,

Your experience as a researcher shows: you raise some subtle but deeply important points about the nature of evaluation. We couldn't agree more that just as we use evaluation to measure the cost-effectiveness of programs, research itself must be a wise and productive use of resources. RCTs are no exception. What we find lacking with your critique is a general point: most of your critiques apply to poorly done research, and are no more or less an issue for RCTs as for any other methodology.

First, let us restate one item. Where RCTs are not appropriate tools for the setting, we are not suggesting to implement them. The tool must fit the question, not the other way around. Having said that, we lack better evidence on what works, in microfinance and elsewhere, and in too many instances randomized trials *could* be done and are not, and would provide better evidence than alternative methods.

Before we get to your criticisms, a critical point needs to come through loudly. Although we are glad to hear that you think that randomized trials address the attribution problem in a “neat way,” the point should be much stronger: it is genuinely better. Methodologies for addressing selection bias can be ordered by their effectiveness at dealing with the problem and RCTs are better than others at doing this. RCTs require fewer assumptions in order to establish causality. The mathematical proof is nicely detailed by Imbens (2009). We encourage a thorough reading of that paper for anyone interested in a more technical explanation of why randomized trials establish causality better than other methods, when feasible. The general principle is simple and two-fold. To establish causality, you need to compare outcomes to *something* so that the analysis is not confounded by outside environmental factors, such as the economy or rainfall. And the comparison group should be similar in terms of observables and unobservables, so that the analyst can remove selection bias (e.g., the entrepreneurial spirit of those who choose to borrow) as a potential explanation for changes in outcomes of those who participate, compared to those who do not. Random assignment tackles both of these problems.

Now let’s go through your points one by one and we will explain how either they are not relevant to RCTs, or that they can be addressed to eliminate the concern.

Point #1 (fitting the tool to the question): You claim that proponents of randomized trials fit their tool to the question, rather than the other way around. No advocate of randomized trials, in our experience, has ever advocated using randomized trials to answer all questions. Furthermore, as we note in the next points, your arguments confuse data collection and random assignment. The random assignment is about putting in place some control to either the offer or delivery of a service. Using RCTs does *not* imply only quantitative data collection! Randomized trials can include ongoing measurement, focus groups, qualitative data collection, etc. And not, as you suggest, only for discrete time periods: RCTs frequently collect data throughout the life of a project to see how impacts develop over time, and to provide intermediate feedback that can then lead to midstream program changes.

Point #2 (external validity): We are glad you raised this because it is the most common and, to be honest, most frustrating critique of RCTs we encounter. It is frustrating because it simply misses the point of what a randomized trial does and does *not* do. You are absolutely right: we do not necessarily know whether a result we find in Peru will hold in Botswana. But truly this has nothing to do with RCTs. RCTs improve the *internal validity* of a study: that is, ensuring that the results from a given study are reliable at that time and place, for that population. The question of external validity—whether results can be generalized elsewhere—is a limitation of *research in general*. It applies to all the research you and we have ever done. At IPA we believe the best way to overcome this is through theory and repetition. By testing ideas in enough places, with enough variety of contexts, we can confidently state that a theory holds and can be used to form policies in new places. We encourage researchers using any methods to do the same.

In your discussion of external validity, you brought up an interesting case about client dropout. This is a nice opportunity to highlight the flexibility and usefulness of RCTs, which often goes unappreciated. Nothing about an RCT suggests you cannot measure dropout, and inquire as to reasons for it. Better yet, an RCT can establish whether product A does better or worse at changing dropout than product B. In Peru, for example, we found that adding business training lowered the dropout rate. So did removing group liability in the Philippines. Is it not better to have concrete evidence of actual decisions made because of an intervention, rather than hoping a small sample of people in a focus group will be accurate in articulating their motivations and predicting future actions?

Naturally one estimate is not what we really need. We need multiple estimates, across choices, to know which is best. RCTs are better as well for pushing us towards making such comparisons. Since they are more internally valid, the comparisons across projects are more feasible and reliable. An ideal chart would list, for example, ten methods of encouraging higher savings so that we can learn, for a dollar spent ten different ways, how much more is saved? Or for a dollar spent on delivering and servicing a loan, given ten different loan structures, which has the biggest impact on client enterprise profits?

Point #3 (costs): We agree, as evaluators we have to be sensitive to the costs of doing research, which can be substantial (but often are not). The biggest mistake is to forget why we are evaluating. RCTs are certainly more expensive than, say, focus groups. But we typically find that non-experimental evaluations cost more, both immediately and in the long run. RCTs are often less costly compared to non-experimental quantitative surveys since they often require smaller sample sizes. In the long run, though, the benefits from better answers as to what works and what does not is worth some investment now. Since results from RCTs are more credible, and therefore more likely to be used to replicate and fund programs found to be effective, they are more cost-effective in many settings.

Point #4 (spillovers): It is key to separate out "natural" spillovers (effects on nonparticipants, e.g., from greater economic activity in the community) from "research" spillovers (a control group individual behaving differently because of awareness that some others got offered a treatment). Three points to make here. First, if designed properly, a study can measure natural spillovers, and RCTs can be effective in helping to measure them with stronger internal validity than other methods. We are currently measuring spillovers on non-borrowers in Mexico and elsewhere. When this can be done, it makes the study that much stronger in its usefulness for setting policy. Second, regarding research spillovers, there are often designs that can take this into account so that it is not a problem. For example, "encouragement" designs typically avoid this by making it acceptable for some in the control group to receive the service if they ask for it. Third, there are situations in which an RCT should not be done for this reason, though typically we have found that creativity can solve the issue.

You asked about the ethics of RCTs. First, the most important “ethical” point to remember is that in most situations there is a scarce resource, and thus randomly allocating versus first-come-first-serve or some other method does not change the number of people who receive a service.

Second, many people think an RCT means you must deny a service to some who eagerly want it. This is simply not true. That is akin to saying you don’t like ice cream because you do not like the flavor chocolate. Encouragement designs, for example, are a common way to conduct a randomized trial when one does not want to deny service to someone who comes forward and asks for it. The design creates a difference between treatment and control groups by doing something (randomly) to encourage some individuals to participate (e.g., marketing or knocking on their door, but not knocking on someone else's door).

Lastly, and perhaps most importantly, we argue that RCTs, when feasible, are *more* ethical. First, in many situations, e.g. when cronyism is possible, random assignment avoids favoritism in the allocation of resources. Second, and most importantly, is it ethical for aid programs to go into developing countries and start projects without clear evidence on what works and what does not? Lives are at stake and funds committed to any one project is money not spent on other noble ideas. We must demand the best evidence we can get to make sure that our current and future decisions help the most people we can. When we take a prescription drug, we demand that it has undergone rigorous testing with randomized trials. Why should we be less demanding for others than we are for ourselves? A dollar spent on microfinance is a dollar not spent on immunizations. Ethics demands that we be as responsible as we can with our money so that we can do the most good possible.

You started by asking about the costs and benefits of randomized trials. We discussed the costs above in point #3, specifically how they are not necessarily more, and often less, than other approaches. An important point remains about the benefits. The benefit of research is driven by the internal and external validity, the relevance of the research question to future decisions, and last but certainly not least importantly, the way the results are communicated and discussed. This Crossfire exchange, we hope, will increase the benefits of randomized trials by dispelling some myths. Your help to set the record straight would do the microfinance community a world of good.

Yours,

Dean and Nathanael

Dear Dean and Nathanael

On ethics, I broadly agree with what you say, though I also note that a lot of medical practice cannot be as easily tested using RCTs as drugs can but that doesn't necessarily mean they are less effective: good research cannot substitute entirely for trusting in the judgement (subject to appropriate safeguards) of professionals.

I have no further quibble over spillover issues, so let me focus on your responses to the other issues I raised.

Problem selection. I'm pleased we agree that RCTs are not appropriate for tackling *all* questions, and hence that using them does not necessarily result in better research. But I'm not fully reassured that we yet have a clear enough picture of which problems can best be tackled by randomization and which not. I appreciate its potential for addressing big design issues, such as the cost-effectiveness of land titling where programme expansion may be close to random anyway. The work in South Africa that found loan ceilings were set too low also demonstrated to me the power of randomized designs. On the other hand, your reference to findings on client exit from your Peru work did not reassure me. You say you found that adding business training in Peru reduced dropout rates. Terrific! But it does not follow that if an MFI has a high exit rate we can most cost-effectively investigate this through a series of RCTs into possible ways of reducing it. If a piece of research yields findings that happen to be relevant to a problem it doesn't follow that it's the best method for addressing the problem. Moving beyond specific examples, we are I think still some way short of having general guidelines on when RCTs work better than other methods and when they don't; and without them subjective preferences will intrude.

External validity. In theory the problem can, as you say, be solved by widespread replication of trials; but how many replications and at what cost? Variation in human bodies is quite small, so good quality trial evidence on whether a drug works on one population is quite likely to be valid for any other. But the way our money management is embedded in history and culture makes me wary of extending the medical metaphor to microfinance. You point out that RCTs can be used to measure how impacts unfold over time, using qualitative as well as quantitative data. Terrific! However, they still seem to require settling on a relatively small number of chosen "treatments" (e.g. ten) to investigate at any one time. Other methods (focus group research into exit rates, for example) can in principle address multiple factors simultaneously and can if necessary be conducted routinely for all clients, if they and their circumstances are diverse enough to warrant it. Such flexibility combined with lower cost and greater divisibility means they can be replicated more widely to address the external validity issue (though you may regard this as primarily a cost issue). I'm drawing here on my experience of agricultural research. Rigorous field trials have been critical to advances in agronomy, but have only been able to tackle a limited range of the constantly changing problems that small-scale farmers confront within diverse, complex, risky and shifting agro-ecological systems. And concentrating resources on them can have an opportunity cost in terms of support for more diversified, flexible and lower cost experimentation by farmers' themselves, whose validity they can often assess reliably enough through their own triangulation against local experience. Hence when references are made to the “credibility” of findings I will continue to ask – for whom, for what purpose and in what context?

Cost effectiveness. I accept there are potential cost advantages of RCTs (especially compared to with-without/before-after impact surveys) and of course I appreciate the benefits of more R&D and piloting prior to mainstreaming. But just as positivist scientific crop research has not eclipsed farmers’ own experimentation so the same applies here. Cost-effectiveness ultimately depends on the external validity issue too, given that heavy costs could result from use of findings without sufficient replication. I assume it is not appropriate to suggest that all MFIs should raise their loan ceilings everywhere, for example? In sum, I look forward to more empirical case studies (including discussion of cost-effectiveness) to corroborate theoretical claims about the potential superiority of RCTS over other methods.

Looking forward to further dialogue

Yours,

James