An Introduction to Impact Evaluations with Randomized Designs[1]

Jonathan **Bauchet**
New York University

Jonathan **Morduch**
New York University

The Financial Access Initiative is a consortium of researchers at New York University, Harvard, Yale and Innovations for Poverty Action.

1. We would like to thank Catherine Burns for her invaluable help reviewing and improving this note. Chapter 9 in Armendáriz & Morduch's *The Economics of Microfinance* (Second Edition, 2010) draws heavily on this note. Kosuke Imai pointed us to useful parts of the statistics literature, and we've drawn on helpful conversations with Rajeev Dehejia. Arguments, interpretations, and errors are ours alone.

FAI IMPACT THROUGH RESEARCH
financial access initiative
www.financialaccess.org

NYU Wagner
Robert F. Wagner Graduate School of Public Service
www.wagner.nyu.edu

1. Introduction

Randomized experiments are increasingly popular ways to evaluate the impacts of development interventions. They provide hope that we can overcome important biases common to nearly all statistical evaluations. When done well, randomized control trials (RCTs) can provide clear, transparent, and credible evidence in complicated contexts, and it's not surprising that they dominate clinical research in medicine.

Still, social science is not medical science, and randomized experiments have limits: they are not always feasible, not always representative, and not always focused on the larger questions of interest.

To see the RCT approach at work, let's say that you offered microfinance services to a group chosen randomly from the population (for example, by applying a random algorithm to select people from a census list) and then selected another group randomly who would be denied access to microfinance. Using the same language as in clinical trials of new pills and medical procedures, the first group is the "treatment" group and the second is the "control" group. The result from statistical theory says that the difference between the average outcome of the treated group and the average outcome of the control group is an accurate estimate of the intervention's average impact. We can interpret the result as the *causal* impact—under certain assumptions, it is a clean estimate of the difference made by microfinance.

Still, social science is not medical science, and randomized experiments have limits: they are not always feasible, not always representative, and not always focused on the larger questions of interest. But already we're seeing their power in studies of the impact of microfinance loans (Karlan and Zinman 2010), savings (Dupas and Robinson 2008, and Ashraf et al. 2006), education interventions (Glewwe et al. 2004), health intervention (Cohen and Dupas 2010, Kremer and Miguel 2004) and many more applications. Below, we describe four examples focused on measuring impacts, one from the Philippines, one on the advantages of access to consumer loans in South Africa, one on microfinance in urban India, and the other on returns to capital of small entrepreneurs in Sri Lanka.

The framing note provides an introduction to impact evaluations with randomized controlled trials. We draw our examples from evaluations of microfinance. We begin by describing the "counterfactual" framework for evaluation, and the biases that arises in typical evaluation contexts. Section 3 explains how randomized designs can overcome these biases, and section 4 describes applications in new studies from the Philippines, India, Sri Lanka and South Africa. In sections 5 and 6 we turn to two important elements for randomized evaluations: the level of randomization, and the need for statistical power. Section 7 focuses on the limits and criticisms of randomized controlled trials, including questions around the ability to generalize to other settings.

Disentangling cause and effect is harder than it might seem at first. The most obvious difficulty is that people can only be in one circumstance at a time.

2. Focus on Causality & Selection Biases

No matter what the outcomes of interest and the intervention are, the most difficult part of evaluating impacts is to separate out the causal role of the intervention. The rough notion of "making a difference" can be translated into a precise question that is at the heart of every credible impact study: How have outcomes changed with the intervention *relative to what would have occurred without the intervention*? The second part of the question is fundamental. In recent decades, education rates and health conditions have improved almost everywhere. Poverty rates too have fallen steadily in a wide range of countries, even where development agencies have had little or no presence. The impact question centers on how an intervention makes a difference over and above these kinds of underlying trends and conditions.

Disentangling cause and effect is harder than it might seem at first. The most obvious difficulty is that people can only be in one circumstance at a time. We can't ever know what would have actually happened to specific individuals had they *not* in fact participated in a development project—just as you can't ever really know what would have happened had you attended a different college, studied different subjects, read different books, or traveled to different places. For example, even if earnings from microfinance participation are funding new houses, further education for children, new savings accounts, and new businesses, we have to ask whether these changes are more remarkable than what would have happened without microfinance. In Banerjee et al. (2009), for example, 69 percent of their baseline sample from urban India had at least one loan outstanding (from moneylenders, family, or friends) *before* microfinance institutions entered the communities. Collins et al. (2009) show how poor households

in developing countries tend to be active managers of their financial lives, using an array of formal and informal savings, loan and insurance products, even without microfinance.

This makes an evaluator's life complicated, since ultimately evaluators want to know whether good outcomes for people might have been nearly as good (or terrible or much better) without the program. To estimate impacts, researchers thus have to find ways to approximate the "counterfactual" (i.e., the prediction of what would have happened without the intervention). Even when it is difficult to form a credible estimate of the counterfactual for a specific individual participant, it can be possible to form a credible estimate for a group of participants taken together. In practice, therefore, the counterfactual is estimated by measuring impacts for individuals who do not participate in the intervention, but are similar to those who do in as many respects as possible.

Observing that microfinance borrowers are wealthier than non-borrowers does not necessarily imply that microfinance made borrowers richer. It could be that being wealthier in the first place made you a better potential borrower...

Establishing a proper counterfactual is at the heart of evaluation methodologies. Two main issues need to be addressed: the "selection" and "reverse causation" biases. The possibility of reverse causation is very real in microfinance. Observing that microfinance borrowers are wealthier than non-borrowers does not necessarily imply that microfinance made borrowers richer. It could be that being wealthier in the first place made you a better potential borrower, so the causal link could run from wealth to microfinance, not the other way around.

Selection bias is the more difficult bias to eliminate. Selection bias arises when individuals are able to self-select into participating in an intervention. In this case, comparing outcomes for participants and non-participants provides an estimate of the impact of the intervention *and* the personal characteristics of the participants that influence the outcomes. In microfinance, individuals who choose to borrow have different personal attributes from those who choose not to borrow. Coleman (2006), for example, reports from Northeast Thailand that households that will later become microfinance borrowers tend to already be significantly wealthier than their nonparticipating neighbors before the microfinance institution starts its operations. It is also likely that households who borrow from microfinance institutions to start or expand a business are more risk-taking than households who prefer not to borrow. In this case, risk tolerance influences both the decision to participate in microfinance and the outcomes that microfinance affects, such as income, wealth or "empowerment." How much of the change in outcome is due to the loan itself and how much is due to the pre-existing characteristics? In the presence of selection bias, it is impossible to tell.

Selection bias can have a very large influence on the impact estimate. In evaluating the Grameen Bank, for example, McKernan (2002) finds that not controlling for selection bias can lead to overestimation of the effect of participation on profits by as much as 100 percent. In other cases, eliminating these biases reverses conclusions about impacts entirely.

To be concrete, consider the impact of microfinance on borrowers' income. Many factors influence a household's income, so identifying the net contribution of the microfinance loans requires a rigorous approach to stripping out selection bias. Some of the individual characteristics that influence both the decision to borrow and income can be observed, measured and controlled in a statistical framework. For instance, gender, age, and education are likely to influence both the decision to borrow and the outcome from having a loan, and this information can be captured by standard surveys of borrowing households.

Some of the individual characteristics that influence both the decision to borrow and income can be observed, measured and controlled in a statistical framework.

The big challenge arises with unobservable factors. Attributes like an individual's entrepreneurial skills, organizational ability, or access to social networks, are far harder—and often impossible—to measure well. But not all hope is lost: randomized designs make it possible to recover the net impact of the intervention, free of selection bias. In randomized experiments, individuals are "assigned" to the treatment and control groups, they do not form the groups themselves. Because an event external to the intervention—a form of lottery—determines who participates in the intervention, the characteristics of the participants are not related to the outcome, and the difference in outcomes between borrowers and non-borrowers is only due to the loan. The next section gives the theoretical justification of this statement.

3. Analytical Foundations of Randomization

Most evaluations compare outcomes for a treatment group, which receives an intervention, and a control group which does not.[2] The outcome for the former can be written as $(Y_1 \mid T)$. In this notation, Y is the outcome and " $\mid T$" means "given that this person received the treatment." The subscript 1 indicates that the outcome Y is measured after having received the treatment. The notation may seem redundant: the subscript 1 and the notation " $\mid T$" appear to refer to the same condition. But, in a subtle and important way, they do not. To see that, first consider a member of the control group. Their outcomes can be written as $(Y_0 \mid C)$. Here, the subscript 0 indicates outcomes without treatment and the notation " $\mid C$" means "conditional on being in the control group." Again, there seems to be a redundancy, this time involving the subscript 0 and the conditioning on C.

2. The treatment here draws on Angrist (2004), Duflo et al. (2008), and Deaton (2009).

As awkward as this notation might seem, it allows us to identify the odd beast, which is the prize of our hunt. This is the term $(Y_1 - Y_0 \mid T)$, which is the difference between the outcome under treatment and the outcome without treatment, for a person in the treatment group: the causal impact. In the case of microfinance, it could be the net effect of access to credit on the profit of an entrepreneur. This is a beast that we don't expect to directly observe in the natural world. We observe $(Y_1 \mid T)$ and $(Y_0 \mid C)$ only, but neither $(Y_0 \mid T)$ nor $(Y_1 \mid C)$. The term $(Y_0 \mid T)$, the expected outcome for an entrepreneur who received a loan *if she had not received that loan,* is not observable. But it is "logically well defined" (Duflo et al. 2008) and the concept helps below.

The key element here is that, for large enough samples, the average attributes of the two groups are comparable before the intervention, because they were formed at random.

Randomizing turns out to yield a simple way to get a handle on $(Y_1 - Y_0 \mid T)$. The term can't be measured for an individual person, but its average value can be measured for a group. The result hinges on the properties of averages. To see that, we introduce the expectations operator and write $E(Y_1 \mid T)$ as the average outcome for all members of the treated group (here, microfinance customers) and write $E(Y_0 \mid C)$ as the average outcome for all members of the control group (Angrist 2004).

So how does one capture $E(Y_1 - Y_0 \mid T)$ from $E(Y_1 \mid T)$ and $E(Y_0 \mid C)$? It turns out that $E(Y_1 - Y_0 \mid T) = E(Y_1 \mid T) - E(Y_0 \mid C)$ if the treatment and controls groups were formed as random samples of the population of interest. They may include residents of villages selected at random from a list of villages, all of which are identified as plausible sites for microfinance expansion. Or they may include interventions targeted to individuals within communities who are chosen at random to receive access to an intervention before their neighbors. The key element here is that, for large enough samples, the average attributes of the two groups are comparable before the intervention, because they were formed at random. If that's so, any differences between the groups after the intervention must be due to the intervention itself.

To see where this result comes from, write:

$$E(Y_1 \mid T) - E(Y_0 \mid C) = E(Y_1 \mid T) - E(Y_0 \mid T) + \{E(Y_0 \mid T) - E(Y_0 \mid C)\} \quad (1)$$

All we've done is subtract and add $E(Y_0 \mid T)$, which is our unobserved hypothetical outcome. Reorganizing the expression by using the fact

that the expectation operator is a linear operator, so the difference of the expectation is the expectation of the difference, we have:[3]

$$E(Y_1 \mid T) - E(Y_0 \mid C) = E(Y_1 - Y_0 \mid T) + \{E(Y_0 \mid T) - E(Y_0 \mid C)\} \quad (2)$$

Our strategy hinges on the term in braces. If it is equal to 0, then $E(Y_1 \mid T) - E(Y_0 \mid C) = E(Y_1 - Y_0 \mid T)$ and we can measure the impact of the loan by comparing the outcomes of treatment and control groups.

The quantity $E(Y_0 \mid T) - E(Y_0 \mid C)$ represents how both the group with credit access and the control group would have fared if nobody had had access. The unobserved beast, $E(Y_0 \mid T) - E(Y_0 \mid C)$, is "selection bias." It is a devil precisely because it is unobservable. This is where the randomization comes into play: if randomization has been completed successfully, this difference is expected to be 0 and vanishes from the expression, leaving us with our prize:

$$E(Y_1 \mid T) - E(Y_0 \mid C) = E(Y_1 - Y_0 \mid T) \quad (3)$$

Randomization promises to banish selection bias, but that pins a lot on the assumption that the randomization has been complete. Without randomizing well, we're back with the troubles that animated section 2 of this note. That's the fear that microentrepreneurs who apply and are approved for loans may well be more dynamic, motivated, risk-tolerant, etc. than microentrepreneurs who do not apply for loans. Or that the locations chosen as sites for microfinance institutions may be particularly promising relative to other sites. "Nonrandom" attrition can also cause problems (say, the least promising customers are the first to drop-out). Contamination of the control group (competitors enter during the study period) is also a worry. In our notation, most of these cases will mean that $E(Y_0 \mid T) > E(Y_0 \mid C)$, biasing upward the estimates of impact. Contamination, or other forms of selection bias, might instead lead to downward biases as E(Y0 | T). Doing randomization well requires that $E(Y_0 \mid T) = E(Y_0 \mid C)$.

One other important note: everything above hinges on the simple properties of expectations of linear operators. That allows us to make claims

The unobserved beast, $E(Y_0 \mid T) - E(Y_0 \mid C)$, is "selection bias." It is a devil precisely because it is unobservable.

3. The fact that "the difference of the expectation is the expectation of the difference" is simply that if, say, you asked a group what their income was last year and you asked them what their income was the year before that, the average change in income for the group could be calculated as either the group's average income change or, equivalently, the group's average income last year minus the group's average income from the year before.

about average impacts. But the basic set-up does not permit us to say anything about the medians and very little about the distributional features of impacts. And we need to be careful in analyzing data on the impacts for particular subgroups in a population. We return to these issues in section 6.

4. Examples of Impact Evaluations with RCTs

Measuring Impacts at the Margin: Consumer Finance in South Africa and Microfinance in the Philippines

Karlan and Zinman (2010) provide an example of a randomized experiment that measures the impact of financial access in South Africa. Here, the institution is not a traditional microlender but a consumer lender that operates commercially and charges high interest rates for short-term (often one month) loans. Unlike most microlenders, the institution tolerates high default rates (loan repayment rates are around 75 percent), and compensates by charging exorbitant interest rates. Still, the study is of interest here since it shows surprisingly positive impacts of consumer lending and demonstrates a creative way to apply randomized methods.

The study design took advantage of the lender's use of credit scoring to allocate loans. In the scoring process, loan applicants are rated on a scale from 100 (most likely to repay) to 0 (least likely to repay). The lender chose a cut-off point below which applicants are excluded from borrowing. The lender, though, feared that the line was too conservative, and the researchers convinced the lender to take a second look at applicants who had narrowly missed being judged creditworthy.

The study focuses on a set of high-risk customers with credit scores in a narrow range just below the cut-off point. From this set, a fraction was chosen (randomly) to be offered a loan. For the lender, the project provided information on the risks and benefits of expanding its approval criteria. For the researchers, the randomization process provided the opportunity to estimate the causal impact of access to the loans. The experiment proceeded by modifying the bank's software. Loan applications were received at the local branch, and loan officers would use proprietary scoring software to evaluate the applicant's creditworthiness. Applicants whose score fell just below the cut-off would normally be denied loans,

but the software was modified to reverse the decision for some of them, chosen randomly. Some marginal applicants would literally have a lucky day. With the process in place, the researchers could investigate average outcomes between the lucky borrowers in the treatment group (325 borrowers) versus the unlucky applicants who were rejected (462 applicants) and thus placed in the control group.[4]

The loans were marketed as consumer loans, but some borrowers used the loans to support microenterprises; most did not. Nonetheless, financial access helped people earn income. Notably, those in the group with access to the loans were more likely to keep their jobs over the study period, which raised their incomes. The median treatment household reported an estimated 16 percent increase in income, and a 19 percent decrease in poverty. Households in the treatment group were 6 percentage points less likely to report that household members had been hungry and 4 percentage points more likely to indicate that food quality had improved in their households since applying for the loan.

The study also showed advantages from the lenders' perspective. First, their credit scoring method proved to have predictive power. The loans approved through the randomization mechanism were indeed less likely to be paid back in full (72 percent for the experimental group versus 76 percent overall). But it also turns out that the additional revenues and costs generated by the experimental loans yielded the lender a net benefit of about US$32 per loan. From the vantage of profit maximization, the credit scoring criteria were too restrictive. In the end, relaxing the lending criteria would be good for client welfare and for the lenders' profits.

From the vantage of profit maximization, the credit scoring criteria were too restrictive. In the end, relaxing the lending criteria would be good for client welfare and for the lenders' profits.

Karlan and Zinman (2009) apply a similar methodology in the Philippines, working again with a commercial lender that made small, uncollateralized loans and charged relatively high interest rates—63 percent when annualized. The institution is First Macro Bank, a for-profit rural bank operating in Metro Manila. This time, however, they targeted low-income microentrepreneurs. Of the 1,601 loan applicants in the sample frame, the credit scoring software randomly approved 1,272 and rejected 329 of them.[5] Researchers conducted follow-up surveys with all of the 1,601 loan applicants. Nearly all of the surveys were completed between one and two years after the individual submitted the loan application.

In this case, the findings were heterogeneous and surprising. Expanding access to credit wasn't associated with an increase in business invest-

4. The researchers measured the impact of the loans on financial access, household welfare, and profitability for the lender. They used administrative data from the lender, credit bureau data about the randomized applicants, and a household survey conducted 6 to 12 months after the start of the experiment (the experiment lasted 2 months, and the loans were standard 4-month loans).

5. The approval rate came from the study's two randomization windows— approve with 60 percent or 85 percent probability. Ultimately, "due to loan officer noncompliance and/or clerical errors," 332 of the approved applicants did not receive a loan and 5 of the rejected applicants did (Karlan and Zinman 2009).

ment, but access was associated with an increase in profit (mostly for men, particularly people with higher income). How did profits rise? Karlan and Zinman (2009) show that members of the treatment group let go of unproductive workers, so their businesses actually shrunk. The results suggest that borrowers used credit to shift business strategies toward smaller, lower-cost, and more profitable businesses. It remains unclear why credit was important in prodding the reoptimization.

The effects on social outcomes in health, education, and women's empowerment were negligible. The study's relatively short time frame, however, limits the scope of the results and their implications to the short-term.

Impact of Microfinance in Urban India

Banerjee et al. (2009) report the first large-scale randomized experiment to measure what happens when microcredit becomes available in a new market. They study 104 similar urban sites in Hyderabad, India. Their baseline survey revealed that there was virtually no formal borrowing in the area prior to the experiment, from microfinance institutions or from commercial banks. About a third of households operated at least one small business, and average profits were 3,040 rupees (about $61).

Spandana, a large microlender, opened branches in 52 of the 104 sites, selected at random. A follow-up survey, conducted at least 12 months after Spandana entered the local market, revealed that households in the treatment areas borrowed almost 50 percent more from microfinance institutions, and were 32 percent more likely to open a business, compared to those in the control areas. Business owners in treatment areas also reported higher profits, but they did not report employing more workers. For households that were already operating businesses at the start of the experiment, investment in durable goods increased significantly. Households identified as likely to start a business (based on characteristics like literacy and the amount of land owned) decreased consumption of nondurable goods such as food and transportation, and of "temptation goods" like alcohol and tobacco in particular. This pattern is consistent with a new entrepreneur's need to make lumpy investments. Households with a low propensity to start a business, on the other hand, increased nondurable consumption. The effects on social outcomes in health, education, and women's empowerment were negligible. The study's relatively short time frame, however, limits the scope of the results and their implications to the short-term. Social outcomes, for example, may take longer to emerge. In the short-run, at least, nothing big and positive leaps out from the evaluation.

Measuring Returns to Capital in Sri Lanka

Suresh de Mel, David McKenzie, and Christopher Woodruff (2008) used another randomized experiment to measure returns to capital for small businesses—a question at the heart of microfinance impacts. Economic theory yields a variety of predictions about returns to capital. One often heard claim flows from the notion of diminishing marginal returns to capital: businesses with less capital are able to produce higher profits per unit of capital than firms with more capital. By this logic, small-scale entrepreneurs should be willing to profit handsomely through microfinance and repay high interest rates. But it is not enough to know that entrepreneurs with access to loans earn high profits since both profits and access to capital depend on "attributes of entrepreneurial ability" (de Mel et al. 2008) and other common causes.

De Mel and his colleagues devised an experiment to introduce randomness in the amount of capital used by businesses. In this way, variation in profits and other outcomes could be pinned on these exogenous increases in capital. The researchers gave some (randomly selected) entrepreneurs larger or smaller grants in cash or equipment/inventory. Randomization guaranteed that the (positive) increase in capital was not correlated with any characteristic of the entrepreneur or its enterprise.

The experiment was based on a survey of small enterprises in Sri Lanka after the tsunami of 2004. The researchers surveyed about 400 firms nine times over a two-year period (2005-2007). The firms were involved in retail sales, manufacturing, or services activities, such as running small grocery stores, sewing clothing, making bamboo products, or repairing bicycles. All firms had US$1,000 or less in capital, excluding land and buildings, at the time of the first survey wave. The grants given to some entrepreneurs were framed as rewards for participating in the survey, to be allocated by a lottery.

Four separate rewards were used, varying by mode of transfer (cash or equipment/inventory) and size of transfer ($100 or $200). If the transfer was in kind, the entrepreneur would get to select their preferred piece of equipment or inventory and it would be purchased by the research team. These transfers were large in relative terms: $100 represented 3 months of the profits generated by the median enterprise, and $200 represented 110 percent of the median firm's capital at the time of the first wave. Cash grants could be used for any purpose, either business- or family-related, and 58 percent of them were actually invested in businesses.

Researchers studied the impact of the capital increase on three outcomes: capital stock, profits, and number of hours worked by the firm's owner. Profits include earnings from the firm's owner, so particular care was taken to estimate the impact on profits net of the impact on the number of hours worked (see de Mel et al. 2009 for a sobering follow-up on measuring profits).

The study showed that the enterprises generated returns to capital ranging from 4.6 to 5.3 percent per month, or about 60 percent per year, depending on the estimation technique. These figures are well above the 16-24 percent nominal interest rates charged by banks and microfinance institutions in the area.

More striking, results indicated considerable heterogeneity in returns. First, the effect for men was large, but no statistically significant average effect was observed for women. (This is an average: some women did well, others poorly.) The finding runs counter to the idea that women are better positioned to take advantage of credit than men, and it aligns with the mixed results in the other studies above. Second, as expected, returns to capital were larger for microenterprise owners with higher ability, as measured by years of schooling and a test of numeracy and cognitive ability. Third, the variation in impacts was very large: half of women entrepreneurs experienced negative returns, and about 20 percent of men had returns lower than the market interest rates. Finally, differences in levels of risk aversion had no discernible impact on returns to capital.

5. Level of Randomization

Some studies randomize at the level of the individual, others randomize treatments across neighborhoods, villages, or another grouping. In microfinance, the options for the unit at which to randomize are most often: the individual, the solidarity group (in group lending contexts), the center, or the branch. In education, one could randomly assign students, classrooms or schools. In many cases, choices are limited by practical constraints. Offering different interest rates to individuals within the same solidarity group, for example, is sure to generate feelings of unfairness within the group. It's often a bad idea for the group, the microfinance institution, and the study.

The choice of unit of analysis is influenced by two important factors: statistical power and the role of spillovers. (For a more advanced discus-

In microfinance, the options for the unit at which to randomize are most often: the individual, the solidarity group (in group lending contexts), the center, or the branch.

sion, see Duflo et al. 2008's excellent toolkit.) When it comes to statistical power, randomizing across groups instead of individuals means that a larger total sample is usually needed to measure the impact of the intervention.[6] Imagine, for example, that villages are assigned to receive a microfinance product or not. To be able to reliably measure effects, the researcher may need to select, say, 100 villages for the treatment group and 100 for the control group. If 20 households are interviewed per village, the total sample would be 4,000 households. If, instead, it was possible to randomize by individuals (so that, within the same village, some people are treated and some people not), the researcher might be able to proceed with just 100 households in the treatment group and 100 in the control—for a sample of just 200 in total. The latter is more appealing in terms of simple costs, but it may not be appropriate or feasible.

The existence of spillovers provides one of the challenges when randomizing at the individual level. Spillovers happen when (i) households transfer from the treatment group to the control group or vice-versa, or (ii) members of the control group are inadvertently affected by the treatment. The second kind of spillover effect can happen, for example, when an entrepreneur receiving a new loan shares some of the loan proceeds with a friend who happens to belong to the control group, or when a microfinance client who receives business training shares some of the lessons and tips with another client who was assigned not to receive the training. Or it could be that, say, improved productivity due to the treatment leads to lower prices in the entire community.

The two forms of spillover affect the random assignment at different levels. Because the identification of impacts relies on the randomness of the assignment to either group, and because individuals rarely switch between treatment and control groups at random, those who switch between groups reintroduce a selection bias in the estimate of impact. The second kind of spillover can reduce (or artificially enlarge) the observed impact of the intervention. For reasons discussed further in the next section, these kinds of spillovers also create a need for a bigger sample. In most cases, some spillovers can be averted by randomizing at the group level rather than the individual or household levels. In a group-lending scheme, for instance, randomly assigning some borrowers inside a group to participate in a program while leaving the others in the control group has a much higher chance of leading to spillovers (and confusion or resentment) than when entire groups are assigned to be either a treatment or control.

Because the identification of impacts relies on the randomness of the assignment to either group, and because individuals rarely switch between treatment and control groups at random, those who switch between groups reintroduce a selection bias in the estimate of impact.

6. We explain statistical power in more details in the next section.

6. Statistical Power

The concept of "power" refers to the ability to reliably detect the impacts of an intervention with statistical methods.[7] Measurement always entails some amount of "noise" due to natural variations in the data and measurement errors. But with a large enough sample, the impact of "noise" can usually be minimized and the effects of interventions emerge clearly. If the sample is too small, the noise may mask the intervention's real effects: measured impacts may be positive and large, but conventional measures of statistical significance would not be able to establish that the measured impacts are nothing other than noise.

Statistical power generally improves with larger sample sizes, but it is not as simple as that. The design of the evaluation matters as well.

This concern is general, but it is more likely with randomized experiments than other approaches because randomized experiments tend to employ smaller samples. "Power" calculations become critical. The calculations illuminate the likely trade-off between detecting the program's effects and keeping sample size in line with research budgets. Statistical power generally improves with larger sample sizes, but it is not as simple as that. The design of the evaluation matters as well.

In our context, the intervention can be microfinance loans, a savings product, a health program offered to microfinance clients, a new program or new loan product that a microfinance institution is thinking about offering, or any similar intervention. Since asking all clients how the intervention affected them is (generally) too costly, a sample of clients is surveyed and statistical methods are used to determine whether conclusions based on the sample can be generalized to all clients. Intuitively, the larger the sample, the more confident one is that findings based on that sample are valid for all clients. The issue is then to make sure that the sample is large enough, but not so large that budgets are busted.[8] This requires a careful balancing act, and the appendix gives a detailed treatment of key issues behind statistical power.

7. Criticisms of Randomization

Randomized experiments have been embraced as the gold standard for evaluations. In many cases they are. But randomization is not always possible, nor always desirable. Lively debates surround claims and counterclaims, and recent views include Deaton (2009), Imbens (2009), Banerjee and Duflo (2009), and Ravallion (2009)—and, from a more technical perspective, Heckman and Smith (1995) and Angrist and Imbens (1994).

7. Duflo et al. (2008) is valuable, and once again we draw from it in this section.

8. This section focuses on how power calculations are used to determine a sample size, pre-study. Power calculations are also used post-study to estimate the level of power obtained with a given sample size.

Many of the criticisms are properly lodged against evaluations in general, not at randomized evaluations specifically. (For example: Are the lessons replicable? Is evaluation worth the trouble and expense?) But some apply to randomization more closely.

First, the randomized methodology provides an estimate of the *average* impact of an intervention. It does not teach us anything about the median impact, and offers little about the distribution of impacts. As illustrated in our power example above, the distribution of the outcome value in the treatment and in the control groups are known, but this does not mean that the distribution of the impact is known.

For example, if a project makes one person much better off and all others a little worse off, a randomized experiment might conclude that the average impact was positive if the positive impact for that one person is large enough to offset the sum of negative impacts for everybody else. Here, the average only catches some of the story. Still, it is not impossible to learn about the distribution of impacts. Building in stratification from the start provides one method. Then impacts can be estimated for subgroups, such as men and women, richer and poorer borrowers, and so on. Consideration of impacts on subgroups ought to be built in from the start, or else the researcher risks "data mining" and finding spurious results. In randomized experiments, as in nonrandomized approaches, specifying in advance which subgroups and hypotheses might be relevant, and restricting one's analysis to these, is key to avoiding data mining. (Yusuf et al. 1991 sound a warning on the temptations of inappropriate subgroup analysis, drawing on medical applications; Assmann et al. 2000 provide a recent survey, also in the medical context.)

Second, while randomized experiments excel at providing a clean estimate of impact, they are by necessity implemented in a particular setting, and may not be easy to generalize to other settings. In technical language, they may have high internal validity but not external validity. A randomized evaluation of flip charts as teacher's aide in schools in Kenya (Glewwe et al. 2004), for instance, only tells us whether the flip charts helped raise test score for these students in these schools in this region of Kenya. One could imagine that students or schools in other parts of Kenya, India, or Latin America have different educational needs, and would benefit differently (or not at all) from flip charts.

Nonrandomized approaches, in contrast, are lauded for making use of data coming from large geographical areas, varied contexts, and/or

diversified populations, so that their conclusions are applicable to a wider range of situations. On the other hand, these methods are often far less satisfactory in terms of internal validity (the question as to whether estimates are credible on their own terms)—and, without that, they don't amount to much.

The limited external validity of randomized experiments takes several dimensions:

1. As highlighted above, randomized experiments are implemented in a specific context, so their results might only apply to that context. Recognizing this limit, proponents of randomized experiments emphasize the need for replications of the experiment in other settings before drawing general conclusions.
2. Because randomized experiments are typically carefully planned and implemented, expansion to a large scale may yield different results. Region-wide policies can seldom be implemented with the same level of care that goes into pilot studies. Still, testing ideas using pilot studies is a smart idea before applying policies on a wide scale. Randomized experiments are well-suited to addressing that need, and they can provide evidence on whether policy ideas really produce measurable impacts on a small scale and under near-ideal conditions.
3. The third issue with external validity has to do with the fact that randomized experiments impose their logic on the operation of the program being evaluated. Absent an experiment, field partners typically do not deny service to a subset of their beneficiaries, and prefer choosing those beneficiaries who have the highest need for, or potential to succeed in, the program. Because randomized experiments require that these two factors be left aside, not all non-government institutions are willing to collaborate with researchers to implement them. If experiments can only be carried out in organizations that accept them, replication will not get rid of the potential selection bias in the choice of field partner. As randomized experiments become more and more common, the hope is that more and more diverse organizations will participate.

Third, randomized experiments follow rigorous designs. In particular, they require that participants respect the initial random assignment to receive the intervention or not—and members of the control group do not, say, pose as members of the treatment group in order to receive benefits.

The advantages of randomization also cease to exist if there are major spillovers between the two groups, and if a nonrandom subset of participants leaves the study. Statistical methods can be employed to correct for spillovers, but at that point the randomness of the assignment has already been undone and experiments have lost some of their edge against nonrandomized approaches.

Fourth, the initial random assignment must be maintained over the course of the study. The problem here is both attrition and contamination. The influence of attrition on the impact estimates is unpredictable, either overestimating or underestimating the impact. Contamination occurs when the organization being evaluated (or another in the same region) either starts working with people in the control areas, or giving added benefits, as a response to the fact that they are not gaining advantages from the treatment.

Fifth, randomized experiments are sometimes criticized on ethical grounds. They indeed require that a portion of the population be denied the intervention that is being evaluated, and the choice of who receives the intervention cannot be made based on fairness considerations ("those who need it most" or "those who deserve it the most"). These concerns are legitimate, and should be taken seriously. In some cases, however, a randomization mechanism may be "fairer" than other selection mechanisms. The selection of beneficiaries of an experimental policy, for instance, or in situations when funding is too limited to serve all eligible individuals, is sometimes fraught with political interventions and favors. Here, publicly randomizing who benefits and who does not can improve the fairness of allocations.

In sum, randomized experiments can be powerful tools to credibly establish that interventions produce impacts. They are not the only method possible, but they have many pluses. Taking their drawbacks seriously as a way to develop improved methods of randomizing and replicating is the next step forward.

In sum, randomized experiments can be powerful tools to credibly establish that interventions produce impacts. They are not the only method possible, but they have many pluses.

References

Angrist, Joshua D., and Guido Imbens. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62 (2): 467–476.

Armendáriz, Beatriz and Jonathan Morduch. 2010. *The Economics of Microfinance*, Second Edition. Cambridge, Mass.: MIT Press.

Ashraf, Nava, Dean Karlan, and Wesley Yin. 2006. "Tying Odysseus to the Mast: Evidence from a Commitment Savings Product in the Philippines." *Quarterly Journal of Economics* 121 (2): 673–697.

Assmann, Susan, Stuart Pocock, Laura E Enos and Linda E Kasten. 2000. "Subgroup Analysis and Other (Mis)uses of Baseline Data in Clinical Trials." *The Lancet* 355 (9209):. 1064-1069.

Banerjee, Abhijit, Esther Duflo, Rachel Glennerster, and Cynthia Kinnan. 2009. "The Micracle of Microfinance? Evidence from a Randomized Evaluation." MIT Department of Economics and Abdul Latif Jameel Poverty Action Lab (J-Pal) Working Paper.

Banerjee, Abhijit V., and Esther Duflo. 2009. "The Experimental Approach to Development Economics." *Annual Review of Economics* (1): 151-178.

Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences*, Second Edition. Hillsdale, NJ: Erlbaum.

Cohen Jessica, and Pascaline Dupas. 2010. "Free Distribution or Cost-Sharing? Evidence from a Malaria Prevention Experiment." *Quaterly Journal of Economics* 125 (1): 1-45..

Coleman, Brett. 2006. "Microfinance in Northeast Thailand: Who Benefits and How Much?" *World Development* 34 (9): 1612–1638.

Collins, Daryl, Jonathan Morduch, Stuart Rutherford, and Orlanda Ruthven. 2009. *Portfolios of the Poor: How the World's Poor Live on $2 a Day*. Princeton, NJ: Princeton University Press.

de Mel, Suresh, David McKenzie, and Christopher M. Woodruff. 2008. "Returns to Capital in Microenterprises: Evidence from a Field Experiment." *Quarterly Journal of Economics* 123 (4): 1329–1372.

de Mel, Suresh, David McKenzie, and Christopher Woodruff. 2009. "Measuring Microenterprise Profits: Must we Ask How the Sausage is Made?" *Journal of Development Economics* 88 (1): 19-31.

Deaton, Angus. 2009. "Instruments of Development: Randomization in the Tropics, and the Search for the Elusive Keys to Economic Development." National Bureau of Economic Research Working Paper 14690.

Duflo, Esther, Rachel Glennerster, and Michael Kremer. 2008. "Using Randomization in Development Economics Research: A Toolkit." in T. Paul Schultz, and John Strauss (eds.), *Handbook of Development Economics*, Vol. 4. Amsterdam: Elsevier Science Ltd., North Holland.

Glewwe, Paul, Michael Kremer, Sylvie Moulin, Eric Zitzewitz. 2004. "Retrospective vs. Prospective Analyses of School Inputs: The Case of Flip Charts in Kenya." *Journal of Development Economics* 74 (1): 251-268.

Hashemi, Syed M. 1997. "Those Left Behind: A Note on Targeting the Hardcore Poor." In Geoffrey Wood and Iffath Sharif (eds.), *Who Needs Credit? Poverty and Finance in Bangladesh*. Dhaka: University Press Ltd.

Heckman, James J., and Jeffrey A. Smith. 1995. "Assessing the Case for Social Experiments." *Journal of Economic Perspectives* 9 (2): 85-110.

Imbens, Guido. 2009. "Better LATE than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)." NBER Working Paper 14896.

Karlan, Dean, Jonathan Morduch and Sendhil Mullainathan. 2010. "Take-up." Financial Access Initiative Framing Note.

Karlan, Dean, and Jonathan Zinman. 2009. "Expanding Microenterprise Credit Access: Using Randomized Supply Decisions to Estimate the Impacts in Manila." Yale University, Dartmouth College, and Innovations for Poverty Action (IPA) Working Paper.

References

Karlan, Dean, and Jonathan Zinman. 2010. "Expanding Credit Access: Using Randomized Supply Decisions to Estimate the Impacts." *Review of Financial Studies* 23 (1): 433-464.

Kremer, Michael and Ted Miguel. 2004. "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities." *Econometrica* 72 (1): 159-217.

McKernan, Signe-Mary. 2002. "The impact of Microcredit Programs on Self-employment Profits: Do Noncredit Program Aspects Matter?" *Review of Economics and Statistics* 84 (1) (February): 93–115.

Ravallion, Martin. 2009. "Should the Randomistas Rule?" *The Economists' Voice* 6 (2): Article 6.

Yusuf, Salim, Janet Wittes, Jeffrey Probstfield, Herman A. Tyroler. 1991. "Analysis and Interpretation of Treatment Effects in Subgroups of Patients in Randomized Clinical Trials." *Journal of the American Medical Association* 266 (1): 93-98.

Appendix: Details on Statistical Power

Power calculations focus on four core elements: (a) the size and variation of the impact, (b) the size of the sample that is used to measure the effect, and (c) two choices about desired levels of statistical significance. The study design matters, so if satisfactory sample and effect sizes cannot be obtained with one design, others should be tried. (We will return to the influence of design elements below.)

Duflo et al. (2008) frame the issue of power in terms of the "minimum detectable effect size" for a given statistical power, significance level, sample size, and study design. The approach is valuable in that it quickly focuses on the trade-off between effect size and sample size. A basic formula for the minimum detectable effect size is

$$MDE = (t_{(1-K)} + t_{\alpha})^* \sqrt{\frac{1}{P(1-P)}} \sqrt{\frac{\sigma^2}{N}} \quad (4)$$

where $t_{(1-K)}$ captures the level of statistical power, t_{α} captures the confidence level, P is the proportion of the sample that receives the treatment, σ^2 is the variance of the effect, and N is the total sample size. Without going into all the details,[9] we reproduce the formula here to highlight the relationship between the minimum detectable effect size and the sample size: as N increases, the minimum detectable effect size decreases, and vice versa. For a given study design, power calculations therefore map the relationship between effect size and sample size, with statistical confidence levels typically kept fixed at 5 percent, 10 percent, and 20 percent.

One practical difficulty is, of course, that the variance of the effect is typically unknown, since the project has not happened yet! The expected effect size is also difficult to determine, which means that it is safe to use as large a sample as possible. Several approaches have been developed to address these issues, some very practical and others more conceptual. The first practical approach is to make a prediction based on previous studies. The second is to do a small pilot study. If neither is possible, estimates are still needed, and it is useful to begin by expressing the effect size in units of the outcome (for example, test scores, dollars of income, number of bed nets used, etc.), or in standard deviations from the mean of the outcome. Cohen (1988) suggests, for example, that an effect of 0.2 standard deviations is small, 0.5 is medium, and 0.8 is large. These numbers, how-

9. A full treatment is available in Duflo et al. (2008) and Bloom (1995, 2005).

ever, need to be placed in the context of the variability of each outcome, and are purely indicative.

The minimum detectable effect size approach and formula also bring to the fore that the relationship between effect size and sample size depends on factors other than the four core elements. First, the proportion of subjects assigned to the treatment and control groups matters. Assigning half of subjects to the treatment group and the other half to the control group makes it possible to detect a smaller effect with a given sample size, or to use a smaller sample to detect a given effect size. We see that since the expression $1/[P^*(1 - P)]$ will be maximized when P = 0.5. If the study involves several treatments groups and one control group, power calculations can indicate the sample size needed for each group.

Second, as we suggested in the previous section, the level of randomization matters greatly for the sample size. The reason is that group-level randomization creates variation between groups, not individuals. Since individuals in a group share some common characteristics, information obtained from each individual brings less variation in the outcome than when the randomization is done at the individual level. Thus, in the former case, more individuals and groups are needed to detect a similar effect size. What matters here is the proportion of the variance in the outcome that comes from the group effect versus that from the individual effect. The higher the former, the bigger the sample needed or the bigger effect size necessary for detection.

Third, some experimental designs do not directly *assign* subjects to treatment and control groups, but "encourage" them to participate in the treatment—say, through an advertising campaign. People in the treatment group can say yes or no to participation, and members of the control group might take up the intervention despite the lack of encouragement directed to them specifically. This design requires a larger sample to achieve the same level of power or detect the same effect size. In their study of microsavings in the Philippines, for example, Ashraf et al. (2006) invited a randomly chosen group of individuals to open a new type of savings account. Some did, some did not. The randomness in this project was in the invitation, not in the opening of an account, so the impacts of the new account must be measured by comparing invited and noninvited individuals. Obviously, not all invited people opened an account. The consequence is that the effect measured at the "invitation level" is diluted and a larger sample size is needed.[10]

10. Karlan et al. (2010) describe issues around the take-up of financial products with an eye to statistical complications.

Finally, well-designed stratified randomized designs can improve the precision of the impact estimate, which makes it possible to use a smaller sample. Stratifying means dividing the sample along one or more observable characteristic, and performing the randomization for each subgroup ("block") separately rather than for the entire sample at once. For instance, stratifying by gender and age could produce four blocks: (1) women over a certain age, (2) women below that age, (3) men over that age, and (4) men below that age. Each block is then assigned to treatment and control. While randomizing individuals into groups create similar groups *in expectation*, stratification is used to ensure that the assignment to treatment or control group is random *in practice* along the dimensions used to stratify. In our example above, if we randomized the same proportion of each block to treatment and control, we know that there will be an equal proportion of each block in the treatment group and an equal proportion of each block in the control group. In effect, stratifying allows analysts to estimate the effect of the intervention for each block separately, although this is done with statistical methods rather than actually repeating the analysis for each block. Because each block is more homogeneous than the entire sample, a smaller variation in outcomes can be detected with the same sample size, allowing for a smaller total sample to be used.