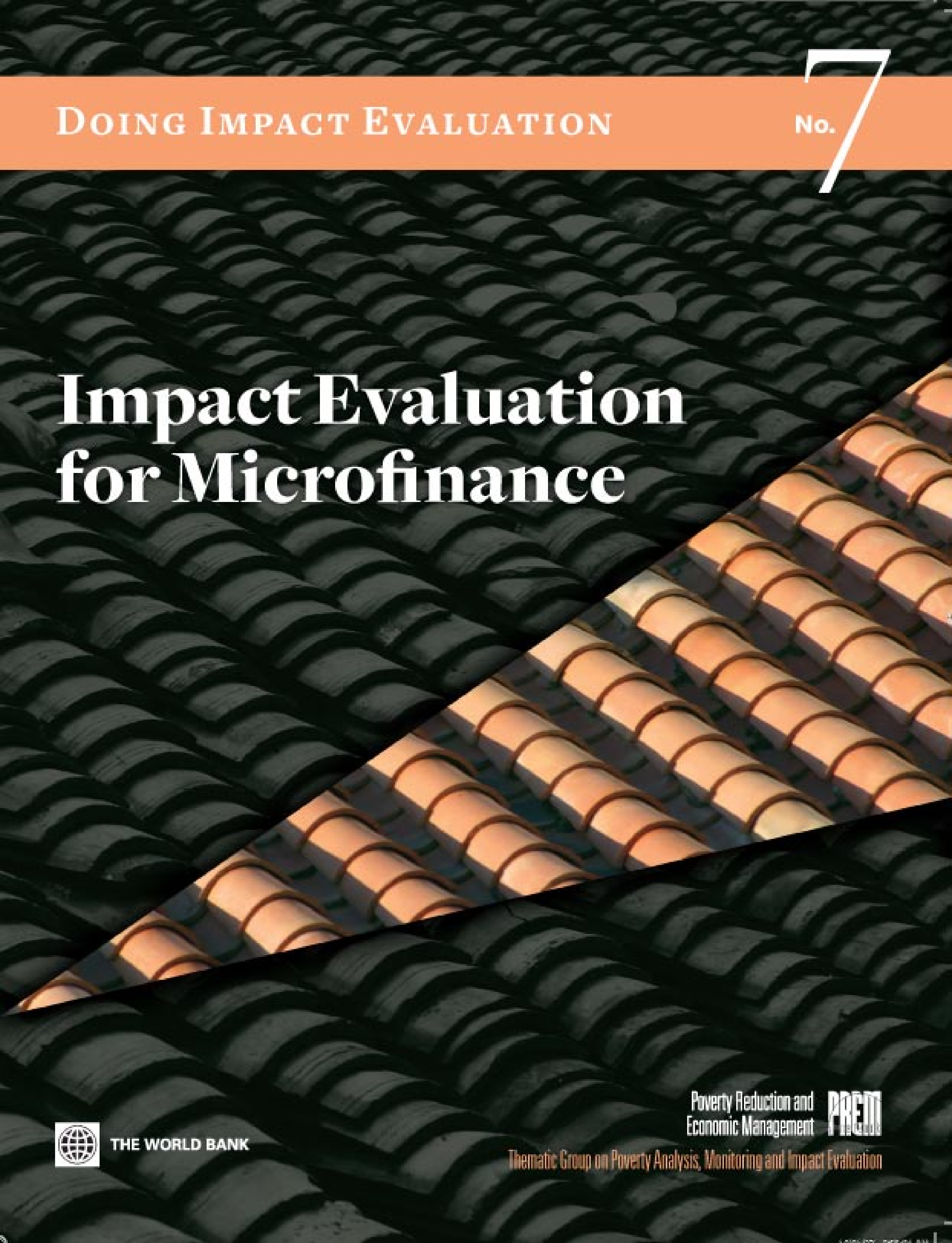


DOING IMPACT EVALUATION No. 7

Impact Evaluation for Microfinance

THE WORLD BANK

Poverty Reduction and Economic Management PREM

Thematic Group on Poverty Analysis, Monitoring and Impact Evaluation

Impact Evaluation for Microfinance: Review of Methodological Issues

November 2007

Acknowledgement

This paper was written by Dean Karlan[1] and Nathanael Goldberg.[2] Comments were provided by Markus Goldstein and Emmanuel Skoufias. This note was task managed by Markus Goldstein and financed by the Trust Fund for Environmentally and Socially Sustainable Development supported by Finland and Norway and by the Bank-Netherlands Partnership Program.

[1] Yale University and Innovations for Poverty Action

[2] Innovations for Poverty Action. The opinions reflected in this paper are the opinions of the authors and not opinions of their institutions.

TABLE OF CONTENTS

Introduction: Why Evaluate?

Impact evaluations can be used either to estimate the impact of an entire program or to evaluate the effect of a new product or policy. In either case, the fundamental evaluation question is the same: “How are the lives of the participants different relative to how they would have been had the program, product, service, or policy not been implemented?” The first part of that question, how are the lives of the participants different, is the easy part. The second part, however, is not. It requires measuring the ***counterfactual***, how their lives *would have been* had the policy *not* been implemented. This is the evaluation challenge. One critical difference between a reliable and unreliable evaluation is how well the design allows the researcher to measure this counterfactual.

Policymakers typically conduct impact evaluations of programs to decide how best to allocate scarce resources. However, since most microfinance institutions (MFIs) aim to be *for-profit* institutions that rely on private investments to finance their activities, some argue that evaluation is unwarranted. At the same time, MFIs, like other businesses, have traditionally focused on quantifying program *outcomes;* in this view, as long as clients repay their loans and take new ones, the program is assumed to be meeting the clients’ needs. Even if this is so, we propose four reasons to evaluate.

First, *an impact evaluation is akin to good market and client research*. By learning more about the impact on clients, one can design better products and processes. Hence, in some cases, an impact evaluation need not even be considered an activity outside the scope of best business practices. For-profit firms can and should invest in learning how best to have a positive impact on their clients. By improving client loyalty and wealth, the institution is likely to keep the clients longer and provide them the resources to use a wider range of services, thus improving profitability. In many cases there also are financial infrastructure investments (e.g., credit bureaus) that improve the market as a whole, not any one particular firm. Public entities may wish to subsidize the research to make sure the knowledge enters into the public domain, so that social welfare is maximized.[3] Note that this point is true both for impact evaluations of an entire program (i.e., testing the impact of expanding access to credit) and impact evaluations of program innovations (e.g., testing the impact of one loan product versus another loan product). We will discuss both types of evaluations in this paper.

Second, *even financially self-sufficient financial institutions often receive indirect subsidies in the form of soft loans or free technical assistance from donor agencies*. Therefore it is reasonable to ask whether these subsidies are justified relative to the next best alternative use of these public funds. Donor agencies have helped create national credit bureaus and worked with governments to adopt sound regulatory policies for microfinance. What is the return on these investments? Impact evaluations allow program managers and policymakers to compare the cost of improving families’ income or health

[3] Note that for-profit firms could have an interest in keeping evaluation results private if they provide a competitive advantage in profitability. However, for-profit firms can and have made excellent socially minded research partners. When public entities fund evaluations with private firms they should have an explicit agreement about the disclosure of the findings.

through microfinance to the cost of achieving the same impact through other interventions. The World Bank's operational policy on financial intermediary lending supports this view, stating that subsidies of poverty reduction programs may be an appropriate use of public funds, providing that they "are economically justified, or can be shown to be the least-cost way of achieving poverty reduction objectives." (World Bank 1998).

Third, *impact evaluations are not simply about measuring whether a given program is having a positive effect on participants*. Impact evaluations provide important information to practitioners and policymakers about the types of products and services that work best for particular types of clients. Exploring why top-performing programs have the impact they do can then help policymakers develop and disseminate best practice policies for MFIs to adopt. Furthermore, impact evaluations allow us to benchmark the performance of different MFIs. In an ideal setting, we would complement impact evaluations with monitoring data so that we could learn which monitoring outcomes, if any, potentially proxy for true impact.

Lastly, *while many microfinance programs aim to be for-profit entities, not all are*. Many are non-profit organizations, and some are government-owned. We need to learn how alternative governance structures influence the impact on clients. Impact may differ because of the programs' designs and organizational efficiencies or because of different targeting and client composition. Regarding the former, many organizations have found they have been better able to grow and attract investment by converting to for-profits. The advantages of commercialization depend on the regulations in each country, and some critics accuse for-profit MFIs of mission drift—earning higher returns by serving better-off clients with larger loans. Some governments have run their own MFIs as government programs. Historically, government-owned programs have had difficulties with repayment (perhaps due to the political difficulty of enforcing loans in bad times), but there are cases where government-owned programs can do well (e.g., Crediamigo in Brazil and BRI in Indonesia).

If, however, the main difference is due to targeting and client composition, impact evaluation is not necessarily needed in the long term. Impact evaluation can begin by measuring the relative impact on the different client pools. However, once the relative impact is known, simpler client profile data and targeting analysis could suffice for making comparative statements across microfinance institutions.

In this paper we seek to provide an overview of impact evaluations of microfinance. We begin in Section I by defining microfinance. This discussion is not merely an exercise in terminology but has immediate implications for how to compare evaluations across different programs. Section II discusses the types of microfinance impacts and policies that can be evaluated, including program evaluation and policy evaluations. Section III reviews experimental and quasi-experimental evaluations methodologies in urban and rural environments and discusses some of the key results from past studies. In Section IV, we review common indicators of impact and sources of data. The final section concludes with a discussion of impact issues that have yet to be adequately addressed.

I. Definition of Microfinance

The first step in conducting an evaluation of a microfinance program is, perhaps surprisingly, to ensure that you are conducting an evaluation of a microfinance program. This seems obvious, but is not, since the definition of "microfinance" is less than clear. Broadly speaking, microfinance for loans (i.e., microcredit) is the provision of small-scale financial services to people who lack access to traditional banking services. The term microfinance usually implies very small loans to low-income clients for self-employment, often with the simultaneous collection of small amounts of savings. How we define "small" and "poor" affects what does and does not constitute microfinance. "Microfinance" by its name clearly is about more than just credit, otherwise we should always call it microcredit. Many programs offer stand-alone savings products, and remittances and insurance are becoming popular innovations in the suite of services offered by financial institutions for the poor. In fact, it is no longer exclusively institutions for the poor that offer microfinance services. Commercial banks and insurance companies are beginning to go downscale to reach new markets, consumer durables companies are targeting the poor with microcredit schemes, and even Wal-Mart is offering remittances services.

Hence, not all programs labeled as "microfinance" will fit everybody's perception of the term, depending on model, target group, and services offered. For example, one recent study collectively refers to programs as varied as rice lenders, buffalo lenders, savings groups, and women's groups as microfinance institutions (Kaboski and Townsend 2005). Another study, Karlan and Zinman (2006b), examines the impact of consumer credit in South Africa that targets *employed* individuals, not micro-entrepreneurs. Surely these are all programs worthy of close examination, but by labeling them as microfinance programs, the researchers are making an implicit statement that they should be benchmarked against other microfinance programs with regard to outreach, impact, and financial self-sufficiency. If the programs do not offer sufficiently similar services to a sufficiently similar target group, it is difficult to infer *why* one program may work better than another. Despite their differences, these programs do typically compete for the same scarce resources from donors and/or investors. Hence, despite their differences and lack of similarities, comparisons are still fruitful since they help decide how to allocate these scarce resources. Note that this argument holds for comparing not only different financial service organizations to each other but also interventions from different sectors, such as education and health, to microfinance. At a macro level, allocations must be made across sectors, not just within sectors. Hence lack of comparability of two organizations' operations and governance structure is not a sufficient argument for failing to compare their relative impacts.

A. Key Characteristics of Microfinance

It may be helpful to enumerate some of the characteristics associated with what is perceived to be "microfinance." There are at least nine traditional features of microfinance:

1. small transactions and minimum balances (whether loans, savings, or insurance),
2. loans for entrepreneurial activity,
3. collateral-free loans,
4. group lending,
5. target poor clients,
6. target female clients,
7. simple application processes,
8. provision of services in underserved communities,
9. market-level interest rates.

It is debatable which of these characteristics, if any, are necessary conditions for a program to be considered microfinance. The first feature, small loans, is likely the most necessary, though lending itself is not essential; some microfinance programs focus on mobilizing savings (although few focus entirely on savings without engaging in any lending). Although MFIs often target microentrepreneurs, they differ as to whether they require this as a condition for a loan. Some MFIs visit borrowers' places of business to verify that loans were used for entrepreneurial activities while other MFIs disburse loans with few questions asked—operating more like consumer credit lenders. In addition, some MFIs require collateral or "collateral substitutes" such as household assets that are valuable to the borrower but less than the value of the loan. Group lending, too, while common practice among MFIs, is certainly not the only method of providing micro-loans. Many MFIs offer individual loans to their established clients and even to first-time borrowers. Grameen Bank, one of the pioneers of the microfinance movement and of the group lending model has since shifted to individual lending.

The focus on "poor" clients is almost universal, with varying definitions of the word "poor." This issue has been made more important recently due to legislation from the United States Congress that requires USAID to restrict funding to programs that focus on the poor. Some argue that microfinance should focus on the "economically active poor," or those just at or below the poverty level (Robinson 2001). Others, on the other hand, suggest that microfinance institutions should try to reach the indigent (Daley-Harris 2005).

Most, but not all, microfinance programs focus on women. Women have been shown to repay their loans more often and to direct a higher share of enterprise proceeds to their families.[4] Worldwide, the Microcredit Summit Campaign reports that 80% of microfinance clients are female. However, the percentage of female clients varies considerably by region, with the highest percentages in Asia, followed by Africa and Latin America, with the fewest women served by microfinance institutions (MFIs) in the Middle East and North Africa. This focus on the poor, and on women, along with the simple application process and the provision of financial services in clients' communities together form financial ***access***, that is, the provision of financial services to the

[4] Higher repayment rates for females is commonly believed but not well documented. In evidence from consumer loans in South Africa (Karlan and Zinman 2006c), women are three percentage points less likely to default on their loans, from a mean of fifteen percent default. Little is known, however, as to why this is so.

unbanked—those who have been excluded from financial services because they are poor, illiterate, or live in rural areas.

Finally, microcredit loans are designed to be offered at market rates of interest such that the MFIs can recover their costs but not so high that they make supernormal profits off the poor. This is an important concept because institutions that charge high interest rates can be scarcely cheaper than the moneylenders they intended to replace, and institutions that charge subsidized rates can distort markets by undercutting other lenders that are attempting to recover their costs. This has implications for impact assessments because the less clients must pay in interest the more they could be expected to show in increased income. If we compare the impact of institutions that fall outside of "normal" microfinance interest rates, we could end up drawing unreasonable conclusions about the effectiveness of one program versus another, since each type of program attracts different clients and imposes different costs on their borrowers.

Note that sustainability of an organization (as defined roughly by *de facto* World Bank policy) does not require each and every product or target market is sustainable but rather that the organization as a whole is sustainable. Thus organizations could charge lower interest rates for indigent or particularly poor individuals, as long as there were sufficient profits from lending to the not-so-poor to be able to cross-subsidize such an program. Such programs may, in the long run, be sustainable (if the initially-subsidized program leads to client loyalty and a long-term relationship with the MFI).

B. Liability Structure of Microfinance Loans

There are three basic models of liability employed by MFIs. Each poses differences in potential impacts (e.g., group-liability programs may generate positive or negative impacts on risk-sharing and social capital) as well as targeting (traditionally, individual-lending programs reach a wealthier clientele).

1. ***Solidarity Groups:*** The classic microfinance model, often referred to as the "Grameen model" after the pioneering Grameen Bank in Bangladesh, involves 5-person solidarity groups, in which each group member guarantees the other members' repayment. If any of the group members fail to repay their loans the other group members must repay for them or they face losing access to future credit.
2. ***Village Banking:*** Village banking expands the solidarity group concept to a larger group of 15-30 women or men who are responsible for managing the loan provided by the MFI (the "external account"), as well as making and collecting loans to and from each other (the "internal account"). In India, Self-Help Groups (SHGs) operate according to a similar format.

Individual Lending: Individual lending is simply the provision of microfinance services to individuals instead of groups. Individual lending can be hard to distinguish from traditional banking since they have similar forms. This is especially true where MFIs require collateral (or collateral substitutes such as household items with low market value but high personal value to the borrower) from borrowers, as collateral-free lending has traditionally been one of the hallmarks of microfinance.

C. "Other" Microfinance Services

Many microfinance programs offer services beyond credit. The most basic such service is savings (credit unions and cooperatives, for instance, rely heavily on savings), although only a few programs focus solely on savings (on the premise that what the poor need most is a safe place to store their money). Some MFIs require mandatory savings each week from each borrower as well as each group, although, depending on the individual MFIs' policies of collection of mandatory savings in case of default, this is often more appropriately called "cash collateral," rather than "savings." Some of these programs also collect voluntary savings, allowing clients to deposit as much as they like each week. Recently MFIs have begun to offer (either independently or bundled with credit) a wide variety of services, including insurance (life insurance and/or health insurance), business development skills training, and remittances. A popular form of training is credit with education, developed by Freedom from Hunger, which includes modules on both business and health training. While MFIs offering credit with education have demonstrated that the modules can be provided at low cost, some MFIs retain their focus on credit and savings, arguing that the poor already have all the business skills they need—what they need most is the cheapest possible source of credit.[5]

II. Types of Policies to Evaluate

We discuss three types of microfinance evaluations: program evaluations, product or process evaluations, and policy evaluations. The World Bank's loans to countries for microfinance typically fit into one of three categories: (1) loans to programs: either loans to state-owned banks that then directly lend to microentrepreneurs (e.g., the Crediamigo program), or loans to second-tier lenders, who then on-lend them to banks (private or public), NGOs or other financial institutions who then on-lend to the poor; (2) technical assistance to help microfinance institutions improve their operations so as to lower costs, expand outreach, and maximize impact; and (3) public policies, such as creating and strengthening credit bureaus, establishing stronger regulatory bodies for savings and capitalization requirements.

The first two of these are the easier ones to evaluate. Public policy initiatives, particularly regulation, are quite difficult to evaluate fully. We will discuss a few examples of when it is possible to learn something about the impact of the policy (such as credit bureaus), but we note that some interventions, particularly those that are implemented at the country level, will be difficult if not impossible to have a full and unbiased evaluation.

[5] See Karlan and Valdivia (2006) for an evaluation of the marginal benefit of business training for microcredit clients. We conduct a randomized control trial in which preexisting credit groups were randomly assigned to either credit with education (business training only) or to credit only (i.e., no change to their services). This random assignment ensures that we are measuring the impact of the business training, and not confounding our result with a selection bias that individual who *want* business training are more likely to improve their businesses regardless of the training. We find that the business training leads to improved client retention, improved client repayment, better business practices, and higher and smoother business revenues.

We divide the types of evaluations into three, and these roughly fit to the above three types of loan purposes. The line between these three is not always crystal clear.

A. First, and perhaps most importantly, ***program evaluation*** refers to examining whether a particular microfinance institution is effective or not in improving the welfare of its clients. In the case of a World Bank loan to a second-tier lender, such an evaluation could be conducted on the institution that receives the money. This is of course not a direct evaluation of the World Bank's loan but rather of the loan received by the institution.

B. Second, ***product*** or ***process evaluation*** refers to evaluating the relative effectiveness for a *particular* microfinance institution in implementing one product versus another, or one process versus another. If the World Bank loan is facilitating technical assistance to microfinance institutions, then here are examples of how evaluations can be done to evaluate not the entirety of the technical assistance, but of particular assistance given on a particular topic. Examples include credit with education versus credit without education, group versus individual liability, and incentive schemes for employees.

C. Third, in the case of ***policy evaluations***, we refer to more macro-level policies, such as regulation of banks and introduction of credit bureaus. Often these macro-level policies do have some micro-level implementation. We put forward examples from interest rate sensitivities to credit bureaus of how to use those micro-level implementations in order to learn the impact of the policy. Some policies, implemented at the macro-level, are arguably not possible to evaluate cleanly. For example, an implementation of new hardware and software for a central bank is undoubtedly outside the scope of an impact evaluation, or changing capitalization requirements for banks may also not be possible to evaluate explicitly.

All three types of evaluations are impact evaluations. Recalling our earlier definition, each of these evaluations distinguishes the outcome from the counterfactual of what would have happened in the absence of the program, process, or policy.

A. Program Impact Evaluations

Historically, MFI impact evaluations have been program evaluations, i.e., they have attempted to measure the overall impact of an MFI on client or community welfare. In many cases, the full package of program services includes many components: credit, education, social capital building, insurance, etc. Thus a program evaluation measures the impact of this full package relative to no package at all. Although useful for measuring whether the resources allocated to the program were worthwhile such program evaluations do not clearly identify which particular aspects of successful programs produced the impact. This type of program evaluation, therefore, will not tell other programs precisely which mechanisms to mimic.

When evaluating the impact of loans to second-tier lenders, though the policy might affect large numbers of people, the evaluation can be pursued in a straightforward manner as a "program" evaluation described above. Loans to banks or MFIs presumably are intended to help the banks or MFIs expand their outreach. By evaluating the impact of such an expansion on client welfare, the multilateral organization providing funding to the second-tier bank can measure the impact of its loan.[6]

B. Product or Process Impact Evaluations

Many microfinance institutions test new product designs by allowing a few volunteer clients to use a new lending product or by offering to a small group of particularly chosen clients (often, their best) a new product. Alternatively, a microfinance institution can implement a change throughout one branch (but for all clients in that branch). We argue that such approaches are risky for lenders, and inferences about the benefits of changes evaluated in such a manner can be misleading. Such approaches do not help establish whether the innovation or change *causes* an improvement for the institution (or the client). Establishing this causal link should be important not only for the microfinance institution implementing the change but also for policymakers and other MFIs that want to know whether they should implement similar changes. This is a situation in which impact evaluations, especially randomized controlled trials, are a win-win proposition: less risky (and hence less costly in the long run) from a business and operations perspective and optimal from a public goods perspective, in that the lessons learned from establishing these causal links can be disseminated to other MFIs.

Examples abound of randomized controlled trials that evaluated the effectiveness for an MFI of a product or process innovation. In each of these cases, the studies measure the impact on the institution. In one study in the Philippines, a bank converted half of its group-liability Grameen-style centers to individual-liability centers. It found that client repayment did not change, client retention improved, and more new clients joined (Giné and Karlan 2006). In South Africa, a consumer finance lender evaluated the sensitivity to interest rates (Karlan and Zinman 2006c; Karlan and Zinman 2006a), as well as the effectiveness of different marketing approaches on the likelihood that individuals borrowed. We find that some costless marketing approaches such as presenting only one rather than several loans or including a woman's photo on the mailer were as effective at increasing demand as dropping the interest rate as much as 4 percentage points per *month* from an average rate across the sample of 7.9 percent (Bertrand, Karlan, Mullainathan, et al. 2005). In Pakistan, in ongoing work, the World Bank, led by Xavier Giné and Ghazala Mansuri, is working with a lender to test different incentive schemes and training for the credit officers, while in India, an ongoing experiment by Erica Field and Rohini Pande through the Center for Microfinance (CMF) is examining the relative merits of different term loans and frequency of payments. In the Philippines, we measured the impact of a new commitment savings product (a specialized savings account with which the client sets a savings goal; her money could not be withdrawn until she reached her goal), as well as an accompanying deposit collection service, and compared the savings balances of clients who receive it to clients who already had traditional savings accounts (Ashraf, Karlan and Yin 2006a; Ashraf, Karlan and Yin 2006b; Ashraf, Karlan and Yin 2006c). In a study in Peru, a village banking organization measured the impact of credit with

[6] Such an evaluation assumes the capital from the loan does not merely crowd out the credit they would have received from other sources.

education to credit without education on both the financial institution (e.g., repayment rates and client retention) as well as client wellbeing (e.g., business profits) (Karlan and Valdivia 2006).

C. Policy Evaluations

Evaluations can also be designed to measure the impact of public policies such as regulatory policies and credit bureaus. Typical regulatory policies include interest rate ceilings and regulation (or prohibition) of savings or savings protection via government deposit insurance programs. It can be difficult to design studies to measure the macro effects resulting from these types of policies. However, there are two ways in which "micro"-level studies can shed insight into the impact of a macro-level policy. First, impacts on specific behaviors in response to policies can be estimated through micro-level interventions that inform individuals about the macro policies. Second, by measuring spillovers on non-participants in micro studies, one can provide community-level estimates of the impacts. This does require typically a large sample, in order to be able to generate variation on the intensity of treatment and then estimate the spillover to non-participants. Depending on the type of spillover, this may or may not be feasible.

An excellent example of the first type of study is recent work in Guatemala on credit bureaus (de Janvry, McIntosh, and Sadoulet 2006). The authors worked with an NGO, Genesis, to assign randomly some clients to receive training on the importance of credit bureaus to their credit opportunities. The clients were informed of both the stick and carrot component (i.e., paying late harms their access to credit elsewhere, yet paying on time gives them access to credit elsewhere at potentially lower rates). The authors find that the training led to higher repayment rates by their clients but also led their clients to borrow elsewhere after establishing a good credit record. This type of study fits under both what we are calling "policy evaluations" as well as "product or process evaluation" (elaborated above). The distinction here is that this particular "process" is intended to help illuminate the effectiveness of the implementation of credit bureaus in Guatemala.

Similar approaches could be applied to a wide variety of policies, such as savings regulation and interest rate policies, and large-scale donor agency initiatives, such as financial infrastructure lending for ATMs, smart cards, and cell phone banking. Such interventions could readily be evaluated with randomized controlled trials of the end products, with treatment groups of participants compared to control groups who do not receive the services.

Regarding savings regulation, two issues in particular seem ripe for evaluation: (1) Do safer, regulated savings, make a difference to individuals when choosing how or whether to save? (2) How does mobilization of savings affect the larger relationship between the MFI and the client? Both of these are consequences of macro-level policies that need to be understood. Naturally, they do not encompass the entirety of the macro-policy and hence should not be seen as a conclusive gross impact of a savings regulatory policy in a country. However, it can provide important information about specific consequences that were generated and can be expected in the future, from approving MFIs to accept savings, to regulating their management of the deposits.

Regarding interest rate policy, two areas should be of particular interest to policymakers and are ripe for carefully executed randomized controlled trials: (1) interest rate caps and (2) consumer protection, ala "Truth in Lending"-type regulation. We have little systematic evidence about sensitivity to interest rates and not much in terms of overall demand nor how different interest rates attract different clients (wealthier vs poorer, riskier versus safer, etc.). Three recent papers from South Africa and Bangladesh demonstrate more sensitivity than is commonly believed (Dehejia, Montgomery and Morduch 2005; Karlan and Zinman 2006c; Karlan and Zinman 2006a). However we do not have enough information, particularly across different countries and settings, to predict confidently what will happen to access to credit if interest rate caps are put in place.[7] Regarding consumer protection, many countries are putting in place laws to regulate how firms present their charges to clients, not just how much they charge. We know there can be tremendous confusion on simple matters of interest. For instance, many lenders charge interest over the declining balance (as is common in developed countries), whereas others charge interest over the initial loan size throughout the life of the loan. Do consumers understand the difference? When given a choice in the market, do they choose the loan that best fits their cash flow needs at the lowest true cost? Studies could be conducted to understand how different presentation of loan terms affects client behavior (take-up, repayment, and impact) in order to then form effective public policies on consumer protection.

III. Methodological Approaches

A. Randomized Controlled Trials for Program Evaluation

Evaluating the impact of a microfinance program requires measuring the impact of receiving the program's services (typically credit, and sometimes savings) versus the counterfactual of not receiving the services. This can be more difficult than evaluating new products or policies (to be discussed below) because the control group must be drawn from non-clients, with whom the MFI does not have a preexisting relationship.

We discuss here three different approaches to conducting experimental evaluations of microcredit programs. In experimental evaluations subjects are selected at the outset, with potential clients randomly assigned to treatment and control groups. When evaluating the impact of an entire program, the treatment as well as the control group must be drawn from potential clients whom the program has yet to serve.

Experimental Credit Scoring

Credit scoring is becoming a popular tool for microfinance institutions seeking to improve the efficiency and speed with which credit is granted (Schreiner 2002). An experimental credit scoring approach uses credit scoring to approve or reject applicants based on their likelihood of default—as with normal credit scoring—but then randomizes

[7] This of course only mentions the demand side of interest rates. Supply side considerations also must be taken into account when formulating interest rate policies.

clients “on the bubble” (those who should neither obviously be approved nor rejected based on the bank’s criteria: e.g., credit history, employment, savings balance) to either receive or not receive credit. The outcomes of those in this middle group who were randomly assigned to receive credit would be compared to those in this middle group who were randomly assigned not to receive credit. The analysis would not examine the outcomes of the clients who fell outside of this randomization “bubble” (i.e., either the extremely good or extremely bad clients). This does have an important implication: the approach measures the impact on only the *marginal* clients with respect to creditworthiness. If access to credit is limited for other reasons (proximity to banking services), this has important implications and may cause an underestimate of the average impact of the program (if those who are most creditworthy accrue more positive benefits from participation) or an overestimate (if those who are least creditworthy accrue more positive benefits from participation). If, on the other hand, the primary contribution of the MFI is that it helps get access to those who are deemed un-creditworthy by other financial institutions, such as commercial banks, then this approach hones in on the exact population of most interest. In other words, perhaps the most creditworthy have other equally good choices for borrowing, hence there is no “impact” (or minimal impact, perhaps) on them and thus measuring the impact on those at the threshold is the exact group that benefits the most.

Note that this approach, if sample sizes permit, does not necessarily require randomization. A *regression discontinuity design* may also be possible if enough individuals are at or near the threshold.[8,9]

The experimental approach also has an operational advantage: it provides lenders with a less risky manner of testing the repayment rates on the marginal (or below marginal) clients. Whereas normally a lender may set a bar at a certain credit score threshold, the randomization allows the lender to lower the bar but limit the number of clients that are allowed in at that level. Furthermore, the experimentation allows the lender to adjust the credit scoring approach. Using a conservative credit scoring approach, which does not allow lenders to test below their normal “approve” level, lenders will never learn whether profit opportunities are being missed because of fear of default.

This approach was employed in a study in South Africa with a consumer lender making micro-loans and is in process with a micro-enterprise lending program in Manila in the Philippines. The lender in South Africa already has a credit scoring system, and the experimental addition focuses strictly on those they normally would reject (whereas the Philippines experiment is as stated above, since no preexisting threshold existed). In South Africa, the lender randomly “unrejects” some clients who had been rejected by the

[8] By comparing a regression discontinuity design to experimental estimates of the PROGRESA program Buddelmeyer and Skoufias (2004) provides useful insight into how far from the discontinuous point one can go without introducing bias into the impact estimate.

[9] The regression discontinuity approach may fail if some individuals near the threshold were given opportunities to improve their application and rise above the threshold.

bank's credit scoring system and branch manager (Karlan and Zinman 2006b).[10] Extending consumer credit to marginal customers produced noticeable benefits for clients, in the form of increased employment and reduced hunger. Plus, the loans to these marginal clients were actually profitable for the lender.

Randomized Program Placement

Randomizing by individual is not always feasible. For example, in implementing a group lending program it would be difficult to go into a rural village and randomly identify individuals to invite to join the group lending program and others who are not invited. Similarly, for a product innovation test, it would be inappropriate to assign randomly some clients to get Credit with Education and others not to when they are in the same lending group, since the classes are given to the group as a whole. Even if you could ask certain clients to remain at the end of the meeting to participate in classes, and others to go home, there may be more subtle reasons why this might not be a good idea. For instance, if you taught a class in marketing skills the treatment group might share the lessons with the control group, or the control group might learn from observing the new marketing techniques of the treatment group. Such spillovers are great if they occur and can be measured. But if simply ignored in the experimental design and measurement, they will then lead to bias in the analysis. The spillovers, when present, are important to measure because doing so not only removes bias from the estimates of the direct effects of the treatment but also provides a measure of the indirect effects. The total program impact is the sum of the direct and indirect effects; thus, when they can be measured it is important for policy purposes to do so. An excellent example of doing just such an exercise is from a deworming intervention of schoolchildren in Kenya (Miguel and Kremer 2004).[11]

In ongoing research in urban India, the Centre for Micro Finance (CMF), the M.I.T. Jameel Poverty Action Lab (J-PAL), and Innovations for Poverty Action (IPA) are

[10] Clients with excessive debt or suspicion of fraud were removed from the sample frame, and all other rejected applicants were randomly assigned credit at a probability correlated with proximity to the approval threshold.

[11] To date, no study has measured such spillovers successfully for microcredit. A few options exist for how to measure them experimentally. Perhaps the best approach would be to randomize the intensity of treatment across geographic areas (e.g., by hiring more credit officers in some areas than others), hence penetrating some areas more than others. Then, one would measure the outcomes on those who do *not* borrow. For example, imagine there are two sets of street markets. In one set, the lender pushes very hard and assigns two credit officers to each market. In the other set, the lender only assigns one credit officer to each market (and assume each credit officer can only handle a fixed number of clients). Are the *non-borrowers* in the two-loan officer street markets worse off than those in the one-loan officer street markets? This would be evidence of negative spillovers, presumably from competitive pressures. Alternatively, if one could collect sufficient baseline information to predict take-up within both treatment and control groups, one could do an experimental propensity score approach and compare the predicted non-borrowers in treatment areas to the predicted non-borrowers in control areas in order to measure the impact on non-borrowers from lending in well-defined geographic areas (e.g., specific markets or rural villages). An alternative approach is to collect detailed data on channels through which impacts flow. This would be most akin to the approach employed in the adoption of agricultural technology literature (Conley and Udry 2005). Note that this can be done in conjunction, or not, with an experimental evaluation (e.g., see Kremer and Miguel (2007)).

conducting an evaluation of the impact of a microfinance program in the slums of Hyderabad. The organization, Spandana, selected 120 slums into which it was willing to expand. The researchers Abhijit Banerjee and Esther Duflo randomly assigned each *slum* to either treatment or control. A baseline survey was completed in each slum, after which Spandana entered the treatment communities and offered loans to as many individuals as possible.[12] At the end of one year, the households from the treatment slums can be compared to the households in the control slums.

A similar design is underway in the Philippines as an extension to earlier work on group versus individual liability (Giné and Karlan 2006). In this study, rural villages in the Philippines are screened by the bank to determine whether they are eligible to be offered lending services. The bank screens to make sure enough micro-entrepreneurs are in the village and express an interest in receiving credit, and to make sure the village chief is amenable to the bank offering its services to the village. The research team then randomizes the eligible villages into four categories: three treatment groups and one control group. The three treatment groups include different lending designs, in order to measure the relative effectiveness of different variations to group and individual liability, while the control group is not offered any lending services. However, the presence of the control group allows the researchers to measure the impact of credit versus no credit on village-level outcomes, as well as individual and microenterprise outcomes.

If randomizing by villages works, it may seem logical to ask: why not randomize by larger units, such as branch or district/area? While such an approach might be good in theory, it greatly limits the number of effective observations in the sample if outcomes are highly correlated within geographic area. It is unusual to come across a setting with a sufficiently large sample size to make it possible in practice. Conversely, simply comparing one branch that gets the treatment to another that does not is not an acceptable strategy. It would be impossible to tell whether the *treatment* worked or whether that branch was different, if perhaps it had an exogenous income shock, such as a particularly good harvest or a new factory generating employment for the region, or had an extraordinarily good (or bad) branch manager.

Encouragement Designs

In ***encouragement*** designs the individuals in the treatment group are *encouraged* to participate in the program (e.g., the program is marketed to them), but they are not required to participate. The program is not marketed to the control group, but they are able to participate if they choose to do so. Therefore, encouragement designs may be useful in situations where it is infeasible to deny service to people who would like to participate in the program.

In encouragement designs, it is critical that ***assignment to treatment*** — as opposed to ***treatment*** — is used to differentiate the groups when analyzing the results. In

[12] Note that for an experimental evaluation, a baseline survey is not necessary. As long as the sample size is large enough, the law of large numbers will produce statistically similar treatment and control groups. Baseline surveys do provide for further statistical precision, as well as the ability to measure heterogeneous treatment effects across more dimensions.

other words, members of the treatment group who do not participate are still part of the treatment group and members of the control group who do participate are still part of the control group. However, it is important to note that the more participating control group members there are, the larger the sample size necessary to detect program impacts. See Duflo and Saez (2004) and Ashraf, Karlan and Yin (2006b; 2006c) for further elaboration and an example of this approach.

Ethical Considerations of Randomized Evaluations

With doubts about the reliability of quasi-experimental designs, randomized evaluations are gaining popularity in international development (Duflo and Kremer 2003). Particularly with poverty programs, however, some observers and policymakers can be uncomfortable with the idea of randomizing the allocation of services to beneficiaries. In instances where the positive benefits of a program seem obvious, the need for an evaluation may come into question. However, until an idea has been properly evaluated, it is wrong to assume that one would be denying the poor a beneficial intervention. It is best to first evaluate the impact and ascertain whether the program does in fact have a positive impact relative to the next-best alternative and then to determine for which types of clients the intervention works best. While microfinance might seem rather benign, there is a very real possibility that taking on debt or paying for services could leave a microfinance client worse off post-intervention. High interest rates are very common in microfinance. But not all clients have the financial sophistication to know their return on investment in their enterprise. Is it possible that the lack of proper recordkeeping causes some clients to continue borrowing (since cash flow increases with the credit and expanded working capital) even though they are actually generating lower profits? Such questions should be kept in mind before one assumes that a given intervention is unambiguously beneficial.

It is important to note that randomized evaluations do not necessarily need to deny services to anybody. A common solution is to randomize the order in which a program expands to an area. Thus, the randomization simply makes use of the organizational constraint that was there in the absence of the evaluation. No fewer people are serviced than before, but by incorporating a random component into the allocation process one generates a clean impact evaluation out of the expansion. Such an approach only works on growing microfinance institutions and ones that are able to plan far enough ahead to generate a list of target areas for a few years. Alternative approaches, such as encouragement designs, are discussed briefly above and in more detail in Duflo, Glennerster, and Kremer (2006).

B. Quasi-experimental Methodologies for Program Evaluation

Quasi-experimental evaluations can be either ***prospective***, in which (as in randomized controlled trials) the treatment group and comparison group are selected in advance of the intervention, or ***retrospective***, in which a comparison group is identified after the intervention. Typically, members of the treatment group are randomly drawn from the MFI’s list of clients. In ***reflexive*** evaluations, or “pre-post,” participants are compared only to themselves before and after the intervention. This is not a useful comparison, however, as many factors could contribute to the changes in their outcomes. For instance, participants' income could increase, but this could be due to general economic changes in the region or simply due to participants acquiring more stable income as they age. In extreme cases, where GDP per capita in a particular country is declining, a reflexive design could show negative impact even if the program succeeded—participants may have fared less poorly than non-participants, hence the program had a positive impact even though participant income fell. We argue that such *reflexive* evaluations should not be referred to as "*impact* evaluations" but rather "client monitoring exercises," or "client tracking exercises," since while they provide information on how clients' lives change, they in no way provide insight into the causal impact of the microfinance program on their lives.

Microfinance evaluators have used a variety of techniques to identify comparison groups. The extent to which these comparison groups adequately mimic the treatment groups is subjective. With microfinance evaluations it can be especially difficult to find a comparison group of similar non-participants, since the non-participants should have the same special (and often unobservable) determination and ability that led the clients to join the microfinance program. Evaluations that compare people clients (those with this special determination) to non-clients will likely overestimate the impact of the programs (assuming this determination, or entrepreneurial spirit, leads to improved business outcomes). The extent to which this increases (or decreases) the estimate of program impact is the ***self-selection bias*** of the non-experimental approach. A related pitfall is bias from ***non-random program placement***, in which outcomes in program villages are compared to outcomes in non-program villages. The problem with this method is that programs choose where they operate for a reason. They may target the poorest villages, for instance, or they may start cautiously with better-off clients before expanding their outreach. The bias from non-random program placement, therefore, can go either way, depending on whether the evaluation compares program villages to non-program villages that may be (even unobservably) better or worse off.

Randomized controlled trials, discussed above, solve these problems. However, it would be a worthwhile exercise to conduct side-by-side experimental and quasi-experimental evaluations and compare the results to determine precisely how far off quasi-experimental evaluations are from experimental evaluations of microfinance programs.[13] If quasi-experimental evaluations can be performed without substantial bias, it will allow evaluators more freedom in their choice of methodology.

Given the potential hazards, it is crucial to ensure that treatment and comparison groups are identical on as many observable dimensions as possible. Comparison group identification techniques have included:

[13] Similar comparisons have been conducted in several settings. LaLonde (1986) finds quasi-experimental evaluations fail to match the results of randomized control trials of labor training programs. Glewwe et al. (2004) finds quasi-experimental evaluations overstate the impact of flip charts in Kenyan schools. Buddelmeyer and Skoufias (2004), however, finds a regression discontinuity design agrees with experimental estimates of the impact of PROGRESA in ten of twelve measures of impact.

- surveying target neighborhoods (either the same neighborhoods in which the treatment groups live or neighborhoods with similar demographics) to identify all households engaged in the informal sector, and then randomly drawing from the list;
- random walk method—starting from a particular point in a neighborhood walking X number of houses to the left, Y number of houses to the right, etc. and attempting to enroll the resulting household in the comparison group.

The quasi-experimental methodology suggested by the USAID-funded project, Assessing the Impact of Microenterprise Services (AIMS), further simplifies the survey methodology by comparing existing clients to incoming clients, suggesting that the difference in outcomes between the two groups represents the impact of the program. Karlan (2001) discusses several flaws with this methodology. Most importantly, if unsuccessful clients drop out, this approach is akin to ignoring one's failures and only measuring one's successes.[14] Furthermore, there may be unobservable reasons why incoming clients differ from clients who chose to enroll in the program at an earlier date. For instance, a year earlier they may have been afraid to join, they may not have had a business opportunity, they may have had a job, or they may have had child-rearing issues. Or, the delay may be due to the MFI. The MFI may not have targeted their village at the time because it was too far from infrastructure like roads and telephones, or because it was too well off. Regardless of the reason, the AIMS-suggested approach will bias the estimate of impact. The punchline often provided to defend this methodology is that "since everyone is a client, they all have entrepreneurial spirit." This argument is flawed. It ignores the time-specific decision to join and assumes that entrepreneurial spirit is a fixed individual characteristic. As the examples above demonstrate, it is easy to imagine that the decision to join a microfinance program is just as much about the time in one's life as it is about the personal fixed characteristics of an individual.

Alexander and Karlan (2006) shows this is not an idle concern. By replicating the AIMS cross-sectional methodology with longitudinal data from one of the AIMS "Core Impact Assessments" of Mibanco, an MFI in Peru, they find several significant differences between existing members and incoming clients, though the directions of the resulting biases differ. New entrants were more likely to have a formal business location, which would understate impact, but were poorer on household measures such as educational expenditures, which would overstate impact.

Coleman (1999) used a novel methodology to control for selection bias; he formed his comparison group out of prospective clients in northern Thailand who signed up a year in advance to participate in two village banks. This technique (later dubbed "pipeline matching") allowed him to compare his estimate of impact to the estimate he would have calculated had he naively compared program participants to a group of non-participants. The "naïve" estimate overstated the gains from participation because

[14] As will be discussed below, clients who exit the program can include both "dropouts" and "successful graduates." The limited evidence available to distinguish between the two types suggests those who exit microfinance programs tend to be worse off on average.

participants turned out to be wealthier than non-participants to begin with. Coleman found no evidence of impact on sales, savings, assets, or school expenditures, and he even found negative effects on medical expenditures and increased borrowing from moneylenders. His results would be more cause for concern, however, if northern Thailand were not already so saturated with credit. 63 percent of the households in the villages surveyed were already members of the Bank for Agriculture and Agricultural Cooperatives (BAAC), a state bank that offered much larger loans than the village banks.

The most ambitious published study to date to control for selection bias and non-random program placement is Pitt and Khandker (1998). Pitt and Khandker, surveying 1,798 households who were members and non-members of three Bangladeshi MFIs (Grameen Bank, BRAC, and RD-12), used the fact that all three programs limited membership to those with landholding totaling less than one-half acre to calculate that every 100 taka lent to a female borrower increased household consumption by 18 taka. Their model ("weighted exogenous sampling maximum likelihood–limited information maximum likelihood–fixed effects") was based on the premise that while there should be no discontinuity in income between people who own just over or just under a half acre of land, participation in the MFIs would be discontinuous because those who were above the cutoff would be rejected from the programs.

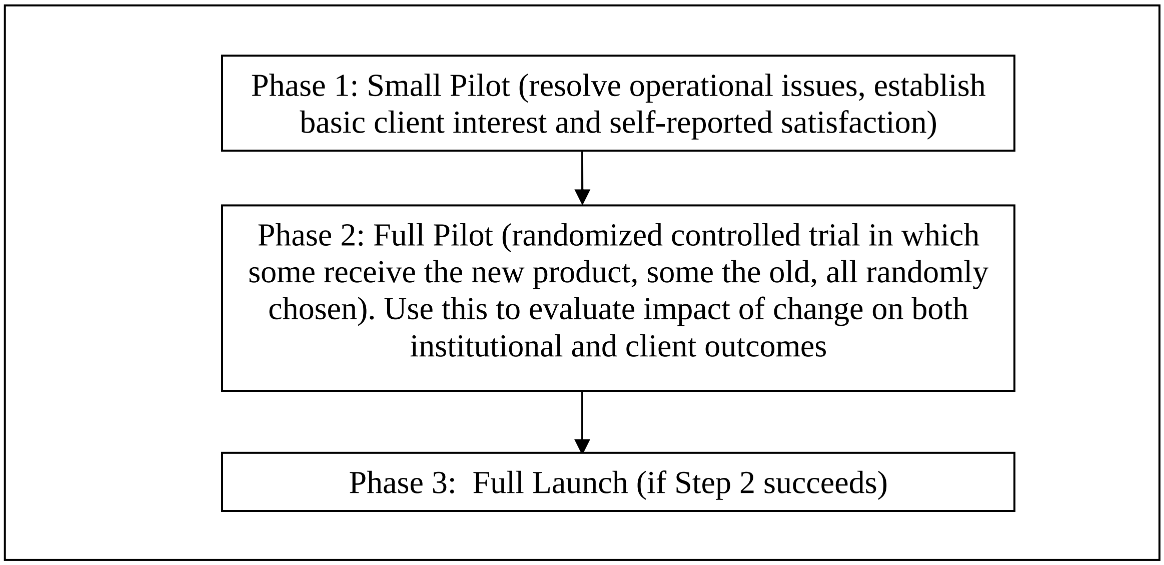
The conclusions we can draw from their findings rely on specific identification assumptions, and the practical implications are also limited in that the methodology is not easily replicated in other settings (and certainly not by practitioners, as it requires involved econometrics). Morduch (1998) challenges the econometric models and identification assumptions in Pitt and Khandker. Using a difference-in-difference model, he finds little evidence for increased consumption but does find reduction in the variance in consumption across seasons.

Khandker (2005) refined their earlier model with the benefit of panel data, which addresses many of the concerns Morduch (1998) raises. In the newer evaluation Khandker finds substantially lower marginal impact on clients. Partially this is because the revised model reduced the estimate of the impact of microfinance, and partially it is because the clients had diminishing marginal returns to credit over time (the first survey was conducted in 1991-2, and the resurvey was performed in 1998-9). In total, Khandker finds an increase in impact—20.5 taka per 100 borrowed versus 18 taka in the earlier paper—because he adds the impacts from both years, 4.2 taka were from current borrowing and 16.3 taka from past borrowing. Poorer clients were found to have larger impacts than the less-poor, and money lent to men was not found to have any impact at all.

C. Randomized Controlled Trials for Product and Process Innovations

In a randomized controlled trial, one program design is compared to another by randomly assigning clients (or potential clients) to either the treatment or the control group. If the program design is an "add-on" or conversion, the design is often simple: The microfinance institution randomly chooses existing clients to be offered the new product. Then, one compares the outcomes of interest for those who are converted to those who remained with the original program. A similar approach is also possible with new clients, although it is slightly more difficult. In this section, we will discuss the logistics of how to change an existing product or process. The following discussion summarizes a process detailed in Giné, Harigaya, Karlan et. al (2006).

The flowchart below presents three basic phases to evaluating the effectiveness of a product or process innovation on the institution and clients. Often, microfinance institutions innovate by doing a small pilot and the full launch (Phases 1 and 3), but not a full pilot (Phase 2). Furthermore, they usually forego random assignment to treatment and control, which would allow them to measure properly the causal link between the product change and institutional and client outcomes. The more common two-stage process involves only a small pilot test to resolve operational issues and gauge interest in the new product and satisfaction among clients who receive it (or sometimes, not even that). If the product "works," the MFI launches the product to all its clients. With this information in hand, the MFI can make much more informed decisions about whether to proceed to a full launch of the innovation and whether to make any changes to the product or policy.


Phase 1: Small Pilot (resolve operational issues, establish basic client interest and self-reported satisfaction)
Phase 2: Full Pilot (randomized controlled trial in which some receive the new product, some the old, all randomly chosen). Use this to evaluate impact of change on both institutional and client outcomes
Phase 3: Full Launch (if Step 2 succeeds)

Product innovation typically aims at solving a problem with the existing product or improving the impact and feasibility of the product. The first step is to identify the problem of the current product and potential solutions through a qualitative process. This should include examination of historical data, focus groups and brainstorming sessions with clients and staff, and ideally discussions with other microfinance institutions that have had similar problems. Once a potential solution is identified, an operating plan and small pilot should be planned. An operating plan should include specifics on all necessary operations components to introduce the proposed change. This includes, for instance, development of training materials, processes for training staff, changes to the internal accounting software, compensation systems, and marketing materials.

In order to resolve operational issues and depending on the complexity of the proposed change, a small pilot implementation should follow. This pre-pilot can be done on a small scale and serves the purpose of testing the operational success of the program design change. Such an endeavor does *not*, however, answer the question of impact to the institution or the client. It instead intends to resolve operational issues so that the full pilot can reflect accurately the true impact.

After the proposed solution has been identified and a small pilot has been conducted, "testing" is not over. The impact of the product innovation on both the institution (repayment rates, client retention rates, operating costs, etc.) and the client (welfare, consumption, income, social capital, etc.) must still be determined. To measure such outcomes properly, one can not merely track the participants and report their changes. One needs a control group.

Often, a proposed solution consists of a main change but many minor issues that need to be decided. For instance, when testing Credit with Education in the FINCA program in Peru (Karlan and Valdivia 2006), the type of education modules to offer had to be selected, and when testing individual liability, the optimal loan size needed to be determined. A careful experimental design can include tests of such sub-questions and collapsed into the evaluation from the start. These questions often arise naturally through the brainstorming questions. Any contentious decision is perfect for such analysis, since if it was contentious then the answer is not obvious.

D. Other Considerations

Determining Sample Size

The minimum necessary sample size depends on the desired effect size (e.g., a 10 percent increase in income), the variance of the outcome, and the tolerance for error in assigning statistical significance to the change in outcome (and the *intra-cluster correlation* if using a clustered randomization, such as randomized program placement). The smaller the minimum detectable difference, the larger the variance, and the lower the tolerance for error, the larger the sample size must be. Outcomes in microfinance evaluations can be both continuous (e.g., change in income) and binary (e.g., no longer below the poverty line). Using binary outcomes can be easier since the variance is entirely determined mathematically from the mean, no data on underlying variation is needed (alternatively, if no variance data are available, one can use standardized effect sizes). Power is weakest for outcomes that have mean 0.50 (the variance is thus 0.25) when the desired effect size is a fixed percentage point increase (e.g., 10 percentage-point increase from 0.5 to 0.6 versus 0.1 to 0.2), but not a percent increase (e.g., a 20 percentage-increase from 0.5 to 0.6 versus 0.1 to 0.12). We recommend the free software Optimal Design to help determine sample sizes, or most statistical packages such as STATA can provide some basic power calculations.[15]

[15] The software can be downloaded from http://www.ssicentral.com/otherproducts/othersoftware.html.

Dropouts

MFIs do not have set lengths of program participation. It is expected that clients will avail themselves of the MFIs' services and leave the programs when they have exhausted the utility of the available products. The more comprehensive the array of products offered, the longer the average client could be expected to "grow" with the program. Broadly speaking, clients who exit an MFI are of two types: those who have outgrown the need for the MFI ("graduates," who hopefully are able to access commercial banking services) and those for whom participation did not bring great benefits ("dropouts"—either they were dissatisfied with the program or they were unable to pay for the MFIs' services).

Without following up with clients it is difficult to distinguish between the two types, and experienced program evaluators understand the importance of including program dropouts in their analysis. Some microfinance evaluation manuals, such as the one offered by AIMS, however, do not counsel evaluators to include dropouts. Alexander and Karlan (2006) demonstrates that failing to include dropouts can bias estimates of impact. They find that after including dropouts some of the measures of impact changed dramatically. Where the AIMS cross-sectional methodology showed an increase of US$1,200 in annual microenterprise profit, including dropouts caused the estimate to fall to a *decrease* of about $170.

In any evaluation, failure to track down enough a sufficiently high percentage of participants can cause ***attrition bias***: if those who cannot be located differ from those who can (it is easy to imagine that this could be the case), the impact estimate can be affected. Those who remain with the program are almost certainly more likely to be located for the follow-up survey than dropouts and more willing to take part in the survey. Not including dropouts at all introduces this problem to an extreme. Whether or not dropouts are less likely to experience a positive impact, if different *types* of clients are more likely to drop out (for instance, richer clients could find it more costly than poorer clients to attend weekly repayment meetings), the composition of the sample will shift and the comparison to the control group will be biased. There are econometric techniques for mitigating these issues.

Targeting

While an impact evaluation is not necessary to evaluate an MFI's outreach to poor clients,[16] when evaluating the impact of a change in program design on existing clients it can be especially useful also to evaluate the impact on the selection process that may result from the change in design (i.e., does the change in program alter the type of client who joins?). There are a couple ways to do this. The simpler method is to compare the demographics of the treatment and control groups, which allows one to say that the change in the program resulted in a different profile of client (e.g., poorer incoming clients) *relative* to the control group. The more powerful method is to conduct (or access)

[16] This can be done with poverty measurement tools on clients and non-clients. For more information see http://www.povertytools.org.

a census survey of households in the treatment and control communities and to compare the distribution of clients in the treatment and control groups to the distribution in the region as a whole. This will allow the MFI to determine the percentage of the population in a given demographic (e.g., below the poverty line) it is currently reaching, as well as the percentage of the demographic it can reach with the new design.

Intensity of Treatment

Intensity of treatment may vary both in length of treatment and quantity of services used. Studies have looked at the impact on clients after one year, two years, even ten years of membership. Deciding at what point to measure impact can be subjective and may depend on the intervention (credit, savings, or another product). There is no set answer but it might be debatable whether one year would be adequate to show impact on credit, for which clients would need time to start /or grow their business. Studies that fail to show impact on one-year clients should acknowledge that the results do not prove that the program has no impact, but merely that it has no impact after one year. The longer the time period, the more difficult it is to employ a randomized controlled trial, since one must maintain the control group throughout the study. Encouragement designs, discussed above, could be useful for longer-term studies as long as the initial "encouragement" has long lasting effects on likelihood of being a client. However, if over time the entire control group gets treated, the encouragement design will fail to measure the long-term impacts as desired. The length of time also relates directly to the outcome measures, as we will discuss in a moment.

IV. Impact Indicators

Microfinance may generate impacts on the client's business, the client's well-being, the client's family, and the community. A thorough impact evaluation will trace the impacts across all of these domains.

In entrepreneurial households money can flow quite easily between the business and different members of the household. Credit is considered fungible, meaning it would be wrong to assume that money lent to a particular household member for a specific purpose will be used only by that person, for that purpose. It is well known, for instance, that loans dispersed for self-employment can often be diverted to more immediate household needs such as food, medicine, and school fees, and that, even though an MFI targets a woman, the loans may often end up transferred to husbands. Thus it would be a mistake to measure only changes in the client's enterprise when evaluating a credit program.

A. Enterprise Income

The most direct outcome of microfinance participation is change in household income and business profits. MFIs almost always work with clients who are engaged in the informal sector and not receiving regular wages. Therefore (as in many developing-country impact evaluations) it can be easier to measure consumption than to measure income.

Business revenue should *not* by itself be considered an impact indicator. Clients who are servicing loans will need to generate increased revenue over and above their loan repayments, or impact will be *negative*, even if business revenue has increased. Therefore, business *profit* is the preferred measure of financial impact on the business. Other business impacts include ownership of business premises and number of employees. Measuring business profits for enterprises without formal records can be difficult. Several options exist, none is perfect. When time permits, it helps to build a flexible survey that allows the surveyor to walk the entrepreneur through their cash flows, starting from their cost of goods sold (or cost of goods produced) per item to revenues per item, and then to frequency of sales. Alternatively, one could focus on funds withdrawn from the enterprise, as well as investments made into the enterprise, in order to back out the net profits. If the family consumes some of the enterprise inventory (as is often the case with buy-sell mini-grocery stores), this approach is more difficult. Similarly, measuring investment in the enterprise can be difficult when inventory levels vary considerably. Hence, this alternative approach should be used cautiously, in settings where business and household lines are kept clearly and when inventory is not highly volatile.

B. Consumption or Income Levels (Poverty)

Evaluations can attempt to determine the number of clients moving out of poverty. This of course requires measuring income (or consumption) versus a standard poverty line. Several studies have developed their own measures of poverty based on a summary statistic of indicators such as housing condition, assets, etc. (Zeller 2005; Schreiner 2006). The World Bank’s Core Welfare Indicator Surveys (CWIQ), which use a reduced set of consumption proxies, could be used in a similar manner. While it may be easier to use such poverty correlates than to measure income, it will limit the reliability of the results and the ability to compare MFIs to other poverty-reduction programs. Depending on the resources available, however, it may be the best alternative. When resources are more plentiful, see Deaton (1997) for more detailed information on proper formulation of consumption surveys. The World Bank Living Standards Measurement Study surveys (LSMS) are also often useful as a starting point for consumption modules in countries around the world. Deaton (1997) discusses many of the advantages and pitfalls of the approaches found in the LSMS.

C. Consumption Smoothing

In addition to changes in income it may also be important to measure the reduction in risk. Many may use credit as an insurance device, helping to absorb negative shocks (Udry 1994). Consumption smoothing can be difficult to measure, since it requires either frequent observations to measure the variance in overall consumption over time, or evidence of particular vulnerabilities. For example, one can measure the number of “hungry days” an individual experienced, or ask about specific negative shocks (illness, death, theft, etc.) and ask how the individual coped with each situation. Although this latter approach is easier in terms of survey complexity, it requires a priori knowledge of the types and sources of risk that the individuals face. If treatment group individuals are better able to cope, this indicates positive impact from access to credit.

D. Wider Impacts

The non-monetary impacts of microfinance participation (i.e., distinct from changes in income) have been labeled "wider impacts." Important examples include children's education and nutrition, housing stock, empowerment, and social capital. While some of these outcomes (e.g., nutrition) can be related to changes in income, others (e.g., women's decision-making power) can be derived from participation in the program itself and the confidence women gain from running a business and handling money. For instance, in the Philippines we find that offering a woman a commitment savings account in her own name leads to an increase in her household decision-making power after one year and that this increase in power leads to more purchases of female-oriented household durables (Ashraf, Karlan, and Yin 2006b). The experimental design for measuring these wider impacts should be much the same as measuring changes in income or poverty, and the data for these outcomes can often be gathered in the same survey. Many of these wider impacts can be measured in a variety of ways, but there may be important differences between indicators that might not be immediately obvious. For instance, height-for-age and weight-for-age (measured in z-scores, or standard deviations) are both measures of malnutrition, but they capture different aspects of severity. Height-for-age ("stunting") is a better indicator of long-term malnutrition, while weight-for-age would better capture acute malnutrition ("wasting").

Other common indicators of nutrition and education include:

- instances per week/month of consumption of specific nutritious foods (e.g., meat, fish, dairy, vegetables) (Husain 1998),
- percentage of children enrolled in school (Pitt and Khandker 1998),
- percentage of possible years of education ("age grade") children have completed (Todd 2001),
- ability to treat children's illnesses such as diarrhea (MkNelly and Dunford 1998),
- medical expenditures (Coleman 1999),
- value of house (Mustafa 1996),
- access to clean water/sanitation (Copestake, Dawson, Fanning et al. 2005)
- use of family planning methods (Steele, Amin, and Naved 1998),
- voted in local or national elections (Cortijo and Kabeer 2004).

E. Spillovers

While it can be simple enough to survey participants and a comparison group of non-participants, restricting our analysis to these groups would misstate the full impact of the program, because the program can be expected to generate impact on non-participants (spillovers) as well. Spillovers can be both positive (increasing community income through increased economic activity) or negative (e.g., if the creation or expansion of participants' enterprises simply transfers sales away from competitors' businesses). This introduces a complication because we do not know every person in the community who will be affected by the program.

In the absence of this information the cleanest method of estimating the true impact of the program is to compare the outcome of entire villages, which can be randomly assigned to treatment or control groups. However, we cannot simply compare participants in the treatment villages to non-participants in control villages because doing so would introduce selection bias—we would be comparing people who chose to join the program to others who did not. Since we do not know who in the control village would have joined the program had it been offered to them we can compare a sample of clients and non-clients in each village to each other. This method measures the impact of ***access*** to microfinance (intent-to-treat effect), rather than ***participation*** in the MFI (treatment on the treated). From a societal perspective, one could argue this is better, as this allows us to reasonably estimate the impact microfinance could have at the macro level. The intent-to-treat effect, since it includes both participants and non-participants in the estimate, will be a lower estimate of expected impact from treating a particular individual, but it can be scaled up by dividing by the probability of participation to obtain the local average treatment effect. The estimate can also be refined with propensity score matching (PSM) if sufficient baseline data are available to predict take-up within the treatment group. This technique re-weights the treatment and control groups by the probability of participating in order to improve the power of the analysis by putting more weight on those more likely to join.

F. Impact on the MFI

When evaluating the effect of new products or policy changes on the MFI the data can usually be collected directly from the MFI's administrative data. Common outcomes of interest for MFIs include the following:

- repayment rate,
- client retention rate,
- new client enrollment,
- average loan size,
- savings balances,
- profitability,
- composition of clients (demographics).

There is a variety of ways to measure the above outcomes. For instance, "profitability" could be financial self-sufficiency, operational self-sufficiency, return on assets, adjusted return on assets, return on equity, and so on. So long as the same definition is used to measure any of the above outcomes before and after the intervention, the chosen definition can serve as a valid indicator of impact. However, the MFI and the microfinance industry may get more value out of the evaluation if standard definitions and financial ratios are used. This way the MFI can measure its performance (and improvement) against others in its peer group. The Microfinance Information Exchange has put forth financial ratio definitions applicable to the microfinance industry.[17]

Several of the impacts on the MFI can be considered "intermediate" indicators, implying that while they are important outputs for the MFI, they do not by themselves indicate a positive outcome for clients. New client enrollment, for example, implies more people have the opportunity to be served by the program, but this will only be a good thing for *clients* if the program improves their welfare, which would be measured through different indicators such as income (described above). Nonetheless, it should be considered a positive indicator for the program, as it has a goal of serving clients.

The World Bank's Poverty Reduction Strategy Paper (PRSP) *Sourcebook* distinguishes between inputs, outputs, and outcomes (Klugman 2002). Inputs and outputs are factors that contribute to achieving outcomes, i.e., impact. Inputs (e.g., funding) contribute to outputs (e.g., number of loans dispersed), and the difference between outputs and outcomes is that outputs are fully under the program's control, whereas outcomes are not. For instance, an MFI can control to whom it disperses loans, but it cannot "create" impact by running clients' businesses for them.

In some cases, the same indicators that measure program outputs can also measure client outcomes. For instance, savings balances are useful to MFIs as a source of loan capital; they are also an indicator of financial stability for clients.

While acknowledging the utility of the distinction between inputs, outputs, and outcomes, we retain the term "impact on the MFI" to indicate the *effect on the input or output* from a change in products or policies. As with impacts on clients, impacts on MFIs need to be measured against a counterfactual of no change.

G. Timing of Measurement

One also should think practically about what types of outcomes are likely to be observed at which points in time. Perhaps the most immediate outcome one should consider is debt level. If the control group has the same quantity as debt as the treatment group, then there is direct evidence that individuals are not credit constrained (the control group simply borrowed elsewhere). This indicates that one should examine the relative *quality* of the debt that each group acquired, since the measurable impact will be driven by difference across debt instruments, not from access versus no access to debt. An intermediate outcome, perhaps six months to one year, would be working capital and/or fixed assets in the business (these may be observable in a shorter time period as well). Increased profits, employment, and formalization may take longer and require one to two years, or more, in which to see the businesses grow sufficiently to observe such impacts. Furthermore, impacts on consumption may be observed immediately if the funds are not used for the enterprise but rather for consumption. If on the other hand the funds are used in the enterprise and profits reinvested, it may take time before the entrepreneur is comfortable withdrawing enterprise funds and increasing consumption.

[17] Available at http://www.mixmbb.org/en/mbb_issues/08/mbb_8.html.

Returning to the discussion at the beginning of this paper, recall that MFIs have often focused on measuring process and institutional measures (e.g., default and client retention) to gauge their performance. However, it is important to note that these types of outcomes may not correlate with client welfare outcomes. In order for MFIs to use these measures as actual impact measures, we must first study whether or not the process and institutional outcomes correlate with client welfare. Such analysis has not been done and would be an important contribution to our knowledge of microfinance.

Conclusion: Outstanding Issues for Evaluation

The microfinance industry needs reliable data, both to prove to donors, governments, and other stakeholders that microfinance works and to improve their products and processes so that they can accelerate their impact on poverty. In the review of the existing impact literature, both from practitioners and academics, Goldberg (2005) finds few if any studies that successfully address the important selection biases relevant for an evaluation of microfinance programs. Randomized controlled trials are the most promising means to allow MFIs to assess reliably the effectiveness of their operations on poverty alleviation and for investors and donors to learn which types of programs produce the strongest welfare improvements. Though several studies are currently underway as discussed earlier, no randomized evaluation of a microfinance program has yet been published.

Evaluations need not be mere costs to an organization in order to prove their worthiness. Quite to the contrary, a good product or process impact evaluation can help an organization improve its operations, maintain or improve its financial sustainability, and simultaneously improve client welfare. The microfinance industry has experienced tremendous experimentation, and now a plethora of approaches exist around the world. How should microfinance institutions decide which approaches to employ when? If evaluation experts worked more closely with microfinance institutions as they make these decisions, we would have better answers and thus prescriptions we could provide to these institutions.

The nine hallmarks of microfinance discussed in the introduction provide a good structure for many of the open questions in microfinance product design:

1. ***Small Loans:*** Certainly microfinance is not microfinance unless loans remain under a certain manageable size, but how small is best for serving the dual needs of the client and the institution? What number of different loan products maximizes impact before becoming unmanageable for the institution and confusing for the client? What other products, such as savings and insurance can be effective complements or substitutes for loans?
2. ***Collateral-free Loans:*** To what extent do collateral requirements or collateral substitutes discourage the poor from participating in MFIs, and to what extent do they raise repayment rates? How effective are collateral substitutes compared to traditional collateral?

3. ***Loans for Entrepreneurial Activity:*** Is this essential for maintaining repayment and ensuring impact on the household? The poor face a variety of credit needs and allowing them to use credit for any type of expenditure could serve them best. Or, loosening the requirement could encourage further indebtedness without a means of escape. To what extent do business skills training help clients manage their enterprises and bolster repayment rates? Why do so many micro-entrepreneurs seem to stagnate at a certain business size, and what can be done to help them expand, employ others, and open additional locations?

4. ***Group Lending:*** Recent evidence from the Philippines and the success of ASA and Grameen II has raised questions about the extent to which high repayments rest on group liability. Can individual liability work as well, or nearly as well?

5. ***Market-level Interest Rates:*** To what extent do high interest rates drive out the poor? Do high rates attract riskier clients? Does subsidized credit "crowd out" market-priced services from competing MFIs?

6. ***Focus on Poor Clients:*** What is the impact of microfinance on the poor? Does microfinance work for the very poor? What specialized services, if any, serve the "poorest of the poor?" Does one need to provide financial literacy along with the loan in order to be effective?

7. ***Simple Application Processes:*** Most MFIs have simple applications, or else they would have few clients. A useful extension is to determine what types of marketing are most effective at increasing take-up of services among the poor.

8. ***Provision of Services in Underserved Communities:*** To what extent does offering credit and savings in poor communities deepen access and increase welfare? Do programs that conduct meetings in the field but require clients to make repayments at the bank branch have lower client retention? Can provision of services in remote areas be profitable?

9. ***Focus on Female Clients:*** Anecdotally, many studies report that women have higher repayment rates than men. Is this true, and if so, what program designs can work best to encourage men to repay their loans? What products and policies can generate the greatest increase in empowerment of female clients?

Impact evaluation of microfinance need not be focused strictly on the impact of credit versus no credit. Instead, prospective evaluation can help MFIs and policymakers design better institutions. Good evaluation not only can deliver to donors an assessment of the benefits that accrued from their investment but also can provide financial institutions with prescriptions for how best to run their businesses and how best to maximize their social impacts.

Bibliography

Alexander-Tedeschi, G. and D. Karlan (2006). "Microfinance Impact: Bias from Dropouts," working paper.

Ashraf, N., D. Karlan, et al. (2006a). "Deposit Collectors," *Advances in Economic Analysis & Policy* 6(2): Article 5.

Ashraf, N., D. Karlan, et al. (2006b). "Female Empowerment: Further Evidence from a Commitment Savings Product in the Philippines," Yale University Economic Growth Center Discussion Paper 939.

Ashraf, N., D. Karlan, et al. (2006c). "Tying Odysseus to the Mast: Evidence from a Commitment Savings Product in the Philippines," *Quarterly Journal of Economics* 121(2): 673-697.

Bertrand, M., D. Karlan, et al. (2005). "What's Psychology Worth? A Field Experiment in Consumer Credit Market," Yale University Economic Growth Center Discussion Paper. 918.

Buddelmeyer, H. and E. Skoufias (2004). "An evaluation of the performance of regression discontinuity design on PROGRESA," World Bank Policy Research Working Paper 3386.

Coleman, B. (1999). "The Impact of Group Lending in Northeast Thailand," *Journal of Development Economics* 60: 105-141.

Conley, T. and C. Udry (2005). "Learning About a New Technology: Pineapple in Ghana," working paper.

Copestake, J., P. Dawson, et al. (2005). "Monitoring Diversity of Poverty Outreach and Impact of Microfinance: A Comparison of Methods Using Data From Peru," *Development Policy Review* 23(6): 703-723.

Cortijo, M.-J. and N. Kabeer (2004). "Direct and Wider Social Impacts of SHARE Microfin Limited: a Case Study from Andhra Pradesh," unpublished Imp-Act report.

Daley-Harris, S. (2005). "State of the Microcredit Summit Campaign Report."

de Janvry, A., C. McIntosh, et al. (2006). "From Private to Public Reputation in Microfinance Lending: An Experiment in Borrower Response," working paper.

Deaton, A. (1997). *The Analysis of Household Surveys*, World Bank.

Dehejia, R., H. Montgomery, et al. (2005). "Do Interest Rates Matter? Credit Demand in the Dhaka Slums," working paper.

Duflo, E., R. Glennerster, et al. (2006). "Using Randomization in Develepment Economics Research: A Toolkit," CEPR working paper 6059.

Duflo, E. and M. Kremer (2003). "Use of Randomization in the Evaluation of Development Effectiveness," paper prepared for the World Bank Operations Evaluation Department (OED) Conference on Evaluation and Development Effectiveness.

Duflo, E. and E. Saez (2004). "Participation in Retirement Benefit Plan," *Quarterly Journal of Economics*.

Giné, X., T. Harigaya, et al. (2006). "Evaluating Microfinance Program Innovation with Randomized Control Trials: An Example from Group versus Individual Lending," working paper.

Giné, X. and D. Karlan (2006). "Group versus Individual Liability: Evidence from a Field Experiment in the Philippines," Yale University Economic Growth Center working paper 940.

Glewwe, P., M. Kremer, et al. (2004). "Retrospective vs. Prospective Analyses of School Inputs: The Case of Flip Charts in Kenya," *Journal of Development Economics* 74: 251-268.

Goldberg, N. (2005). "Measuring the Impact of Microfinance: Taking Stock of What We Know," Grameen Foundation USA publication series.

Husain, A. M. M. (1998). "Poverty Alleviation and Empowerment: The Second Impact Assessment Study of BRAC’s Rural Development Programme," BRAC publication.

Kaboski, J. and R. Townsend (2005). "Policies and Impact: An Analysis of Village-Level Microfinance Institutions," *Journal of the European Economic Association* 3(1): 1-50.

Karlan, D. (2001). "Microfinance Impact Assessments: The Perils of Using New Members as a Control Group," *Journal of Microfinance*.

Karlan, D. and M. Valdivia (2006). "Teaching Entrepreneurship: Impact of Business Training on Microfinance Institutions and Clients," Yale University Economic Growth Center working paper.

Karlan, D. and J. Zinman (2006a). "Credit Elasticities in Less Developed Economies: Implications for Microfinance," working paper.

Karlan, D. and J. Zinman (2006b). "Expanding Credit Access: Using Randomized Supply Decisions to Estimate the Impacts," working paper.

Karlan, D. and J. Zinman (2006c). "Observing Unobservables: Identifying Information Asymmetries with a Consumer Credit Field Experiment," working paper.

Khandker, S. R. (2005). "Micro-finance and Poverty: Evidence Using Panel Data from Bangladesh," *World Bank Economic Review* 19(2): 263-286.

Klugman, J. (2002). *A Sourcebook for Poverty Reduction Strategies* World Bank.

Kremer, M. and E. A. Miguel (2007). "The Illusion of Sustainability," *Quarterly Journal of Economics*, forthcoming.

LaLonde, R. J. (1986). "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *American Economic Review* 76(4): 604-620.

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

MkNelly, B. and C. Dunford (1998). "Impact of Credit with Education on Mothers and Their Young Children's Nutrition: Lower Pra Rural Bank Credit with Education Program in Ghana," Freedom from Hunger publication.

Morduch, J. (1998). Does Microfinance Really Help the Poor? New Evidence on Flagship Programs in Bangladesh," MacArthur Foundation project on inequality. Princeton University, draft.

Mustafa, S. (1996). "Beacon of Hope an impact assessment study of BRAC's Rural Development Programme," BRAC publication.

Pitt, M. and S. Khandker (1998). "The Impact of Group-Based Credit Programs on Poor Households in Bangladesh: Does the Gender of Participants Matter?" *The Journal of Political Economy* 106(5): 958-996.

Robinson, M. S. (2001). *The Microfinance Revolution*, The World Bank and Open Society Institute.

Schreiner, M. (2002). "Scoring: The Next Breakthrough in Microcredit?" working paper.

Schreiner, M. (2006). "Seven Extremely Simple Poverty Scorecards," working paper.

Steele, F., S. Amin, et al. (1998). "The Impact of an Integrated Microcredit Program on Women's Empowerment and Fertility Behavior in Rural Bangladesh," Population Council publication.

Todd, H. (2001). "Paths out of Poverty: The Impact of SHARE Microfin Limited in Andhra Pradesh, India," unpublished Imp-Act report.

Udry, C. (1994). "Risk and Insurance in a Rural Credit Market: An Empirical Investigation in Northern Nigeria," *Review of Economic Studies* 61(3): 495-526.

World Bank (1998). *World Bank Operational Manual OP 8.30 - Financial Intermediary Lending*.

Zeller, M. (2005). "Developing and Testing Poverty Assessment Tools: Results from Accuracy Tests in Peru," IRIS working paper.